

Differentially Private Data Synthesis Methods

Claire McKay Bowen^{*} and Fang Liu^{† *}

Abstract

When sharing data among researchers or releasing data for public use, there is a risk of exposing sensitive information of individuals who contribute to the data. Data synthesis (DS) is a statistical disclosure limitation technique for releasing synthetic data sets with pseudo individual records. Traditional DS techniques often rely on strong assumptions on a data intruder's behaviors and background knowledge to assess disclosure risk. Differential privacy formulates a theoretical approach for strong and robust privacy guarantee in data release without having to model intruders' behaviors. In recent years, efforts have been made aiming to incorporate the DP concept in the DS process. In this paper, we examine current **D**ifferentially **P**rivate **D**ata **S**ynthesis (dips) techniques, compare the techniques conceptually, and evaluate the statistical utility and inferential properties of the synthetic data via each dips technique through extensive simulation studies. The comparisons and simulation results shed light on the practical feasibility and utility of the various dips approaches, and suggest future research directions for dips.

keywords: differential privacy, dips, Laplace sanitizer, perturbed histogram, smooth histogram, statistical disclosure limitation

^{*}Claire M. Bowen is a graduate student, and Fang Liu is a Huisking Foundation, Inc. Assistant Professor in the Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556 ([†]E-mail: fang.liu.131@nd.edu). Claire Bowen is supported by the National Science Foundation (NSF) Graduate Research Fellowship under Grant No. DGE-1313583. Fang Liu is supported by the NSF Grant 1546373 and University of Notre Dame Faculty Research Support Initiation Grant Program.

1 Introduction

When sharing data among collaborators or releasing data publicly, one big concern is the risk of exposing personal information of individuals who contribute to the data. Even with key identifiers removed, a data intruder may still identify a participant in a data set such as using linkage with public information. Some notable examples on individual identification breach in publicly released or restricted access data include the Netflix prize [[Narayanan and Shmatikov, 2008](#)], the genotype and HapMap linkage effort [[Homer et al., 2008](#)], the AOL search log release [[Götz et al., 2012](#)], and the Washington State health record identification [[Sweeney, 2013](#)].

Statistical approaches to protecting data privacy are referred to as statistical disclosure limitation (SDL) in general. SDL techniques aim to provide protection for sensitive information while releasing individual-level data for research and public use. Data synthesis (DS) is a SDL technique that focuses on releasing individual-level data synthesized or imputed from an appropriate model constructed based on the original data [[Drechsler, 2011](#)]. One DS approach is replacing a portion or the full original data set with its perturbed values (such as the posterior predictive values from a Bayesian statistical model). To propagate the uncertainty arising from the synthesis process, multiple synthetic sets of the identical structure are often released. This procedure is referred to as multiple synthesis (MS). Methods have been developed to combine the results from multiple synthetic data sets to yield valid final inferences [[Raghunathan et al., 2003](#), [Reiter, 2002, 2003, 2005](#)].

Existing disclosure risk assessment approaches in SDL techniques often depend on the specific values in a given data set, as well as various assumptions about the background knowledge and behaviors of data intruders [[Hundepool et al., 2012](#), [Manrique-Vallier and Reiter, 2012](#)]. In some cases, only

heuristic arguments are employed without numerical assessment of disclosure risk, such as “there is no disclosure risk in the released data since no records correspond to any real persons,” which could be an over-optimistic statement [Abowd and Vilhuber, 2008].

Differential privacy (DP), a concept popularized in the theoretical computer science community, provides strong privacy guarantee in mathematical terms without making assumptions about the background knowledge of data intruders [Dwork et al., 2006, Dwork, 2008, 2011]. DP was originally developed for releasing aggregate or summary statistics to queries submitted to a database. This method is known as the interactive privacy mechanism or the query-based method. For a given privacy budget, the released query results leak no additional personal information of an individual from the data nor can be inferred. This privacy guarantee holds regardless of how much background information the data user possesses about the individual. DP has spurred a great amount work in developing differentially private mechanisms in general settings [Dwork et al., 2006, McSherry and Talwar, 2007, McSherry, 2009], as well as for specific statistical analysis such as contingency table [Barak et al., 2007], robust statistical estimators [Dwork, 2011], efficient point estimators [Dwork and Smith, 2010], machine learning [Blum et al., 2008], data mining [Mohammed et al., 2011], shrinkage regression [Chaudhuri et al., 2011, Kifer et al., 2012], principle component analysis [Chaudhuri et al., 2012], and genetic association tests [Yu et al., 2014].

However, query-based data release has several shortcomings. The requirement to pre-specify the level of privacy budget ϵ often dictates the number and the types of future queries. The data curator of a database will refuse to answer any further queries if the pre-specified privacy budget ϵ is exhausted from answering all previous queries. From the perspective of data users, they would prefer to have access to individual-level data rather than sending queries and receiving perturbed query results from

statistical analysis not done by themselves. Efforts have also been made to release differentially private individual-level data sets. We refer to this type of approach as dips (**D**ifferentially **P**riate **D**ata **S**ynthesis). [Abowd and Vilhuber \[2008\]](#) proposed an approach to synthesize differentially private tabular data from the predictive posterior distributions of frequencies. The same technique was applied in the simulations studies in [Charest \[2010\]](#) to explore the statistical inferences on proportions from synthesized binary data. [McClure and Reiter \[2012\]](#) proposed synthesizing differentially private data by sampling from the posterior distribution of the cell probabilities of a data set. [Blum et al. \[2008\]](#) suggested generating a synthetic data set using a machine learning approach. [Wasserman and Zhou \[2010\]](#) proposed three paradigms to sample from appropriately differentially private perturbed histograms or empirical distribution functions, and examined the rate at which the probability of empirical distribution of the synthetic data converges to the true distribution of the original data. [Liu \[2016a\]](#) proposed a Bayesian technique, model-based dips (modips), to release model-based differentially private synthetic data, and explored the inferential properties of the released data. This paper aims to compare the currently available dips approaches both conceptually and empirically via simulation studies, and provide guidance on the feasibility of the dips methods for practical use.

The remainder of the paper is organized as follows. [Section 2](#) overviews the basic concepts in differential privacy and two common differentially private release mechanisms. [Section 3](#) presents the non-parametric dips (np-dips) and the model-based dips approaches. [Section 4](#) compares and examines the utility and the inferential properties of the data released from the dips methods introduced in [Section 3](#) via three simulation studies. Concluding remarks are given in [Section 5](#).

2 Concepts

The original concepts of differential privacy and the sanitization algorithms were developed for releasing query results sent to a database. We present the concepts below in terms of statistics. There is essentially no difference between query results and statistics given that the queries are also functions of data, but statistics are more compatible with the data synthesis process presented in Section 3.

Denote the target data for protection by $\mathbf{x} = \{x_{ij}\}$ of dimension $n \times p$. Each row \mathbf{x}_i ($i = 1, \dots, n$) represents an individual record with p variables/attributes ($j = 1, \dots, p$). One key assumption underlying the following discussions is that every variable in \mathbf{x} data is bounded, the reason being that it is difficult to extend differential privacy to unbounded domains [Sarathy and Muralidhar, 2009, Wasserman and Zhou, 2010]. This assumption is generally true in real life settings. Categorical data is “bounded” because every categorical variable can be coded with binary dummy variables. Real-life numerical variables are hardly unbounded. For example, it is safe to say human height is bounded within $(0, 300)$ cm, and personal annual income is bounded within $[0, 10]$ billion.

2.1 Differential privacy

Differential privacy (DP) was first proposed by Dwork et al. [2006] and is rephrased in terms of statistics $\mathbf{s}(\mathbf{x})$ in Definition 2.1.

Definition 2.1. A sanitization/perturbation algorithm \mathcal{R} gives ϵ -differential privacy if for all data sets $(\mathbf{x}, \mathbf{x}')$ that is $\delta(\mathbf{x}, \mathbf{x}') = 1$, and all results $Q \subseteq \mathcal{T}$

$$\left| \log \left(\frac{\Pr(\mathcal{R}(\mathbf{s}(\mathbf{x})) \in Q)}{\Pr(\mathcal{R}(\mathbf{s}(\mathbf{x}')) \in Q)} \right) \right| \leq \epsilon, \quad (1)$$

where \mathcal{T} denotes the output range of the algorithm \mathcal{R} , and $\epsilon > 0$ is the privacy “budget” parameter.

$\delta(\mathbf{x}, \mathbf{x}') = 1$ implies that \mathbf{x}' differs from \mathbf{x} by only one individual. Mathematically, Equation (1) states that the probability of obtaining the same query result to a query sent to sanitized \mathbf{x} and \mathbf{x}' is roughly the same. In layman's terms, DP means the chance an individual will be identified based on the sanitized query result is very low since the query result would be about the same with or without the individual in the database. The degree of "roughly the same" is determined by the value of ϵ . The smaller the privacy budget, ϵ , the probabilities of obtaining the same queries from \mathbf{x} and \mathbf{x}' will become more similar. DP provides a strong and robust privacy guarantee in the sense that it does not assume anything regarding the background knowledge or the behavior of a data intruder.

2.2 Differential private mechanisms

There are two commonly used data release mechanisms to achieve ϵ -DP for the release of \mathbf{s} : the Laplace mechanism [Dwork et al., 2006] and the Exponential mechanism [McSherry and Talwar, 2007]. A key concept in the Laplace mechanism is the global sensitivity (GS) of \mathbf{s} [Dwork et al., 2006], defined as the following: For all $(\mathbf{x}, \mathbf{x}')$ that is $\delta(\mathbf{x}, \mathbf{x}') = 1$, the global sensitivity of \mathbf{s} is $\delta_{\mathbf{s}} = \max_{\mathbf{x}, \mathbf{x}', \delta(\mathbf{x}, \mathbf{x}')=1} \|\mathbf{s}(\mathbf{x}) - \mathbf{s}(\mathbf{x}')\|_1$. In layman's terms, $\delta_{\mathbf{s}}$ is the maximum change you would expect in \mathbf{s} across all possible configuration of \mathbf{x}, \mathbf{x}' , and $\delta(\mathbf{x}, \mathbf{x}') = 1$. The sensitivity is "global" since it is defined for all possible data sets and all possible ways that two data sets differ by one row. The higher $\delta_{\mathbf{s}}$ is, the more disclosure risk there will be on the individuals in the data from releasing the original \mathbf{s} .

Definition 2.2. The Laplace mechanism adds an independent noise term from the Laplace distribution with rate parameter $\delta_{\mathbf{s}}\epsilon^{-1}$ to each of the elements of the original result \mathbf{s} . The perturbed result $\mathbf{s}^* = \mathbf{s} + \mathbf{e}$ satisfies ϵ -DP.

By the Laplace distribution, values closer to the raw results \mathbf{s} have higher probabilities of being released

than those that are further away from \mathbf{s} . The variance of the Laplace distribution is $2(\delta_{\mathbf{s}}\epsilon^{-1})^2$, implying the smaller the privacy budget ϵ and/or the larger the $\delta_{\mathbf{s}}$, the higher the likelihood that \mathbf{s}^* will be farther way from \mathbf{s} when released.

Definition 2.3. In the Exponential mechanism, a scoring function u assigns a score to each possible output \mathbf{s}^* and releases \mathbf{s}^* with probability

$$\frac{\exp\left(u(\mathbf{s}^*|\mathbf{x})\frac{\epsilon}{2\delta_u}\right)}{\int \exp\left(u(\mathbf{s}^*|\mathbf{x})\frac{\epsilon}{2\delta_u}\right) d\mathbf{s}^*} \quad (2)$$

to ensure ϵ -DP, where $\delta_u = \max_{\mathbf{x}, \mathbf{x}', \delta(\mathbf{x}, \mathbf{x}')=1} |u(\mathbf{s}^*|\mathbf{x}) - u(\mathbf{s}^*|\mathbf{x}')|$ is the maximum change in score u with one row change in the data.

The Exponential mechanism was introduced by [McSherry and Talwar \[2007\]](#). Per the Exponential mechanism, the probability of returning \mathbf{s}^* increases exponentially with the utility score. There are many ways to define the utility function, depending on the goal of releasing data. If the goal is to preserve as much original information as possible, which is often the case when releasing individual-level data for statistical analysis, metrics measuring the closeness between \mathbf{s}^* and the original \mathbf{s} are good candidates for u , such as the negative p -norm distance $-||\mathbf{s} - \mathbf{s}^*||_p = -\left(\sum_{j=1}^r |s_j - s_j^*|^p\right)^{1/p}$ between \mathbf{s} and \mathbf{s}^* [\[Liu, 2016b\]](#). The closer \mathbf{s}^* is to \mathbf{s} , the larger the utility score $u(\mathbf{s}^*|\mathbf{x})$ is, and the higher the probability the corresponding \mathbf{s}^* will be released. When the L_1 norm distance is used, the Exponential mechanism in Definition (2) becomes the Laplace mechanism with halved privacy budget [\[McSherry and Talwar, 2007, Liu, 2016b\]](#).

3 Differentially private data synthesis (dips)

We group the currently available dips methods into two categories: the non-parametric approach (np-dips) and the model-based approach (modips). In the np-dips approach, the synthesizer is constructed

based on the empirical distributions and density histograms of \mathbf{x} [Abowd and Vilhuber, 2008, Wasserman and Zhou, 2010]. In the modips approach, the synthesizer is built upon an appropriately defined Bayesian model given \mathbf{x} and correctly identified Bayesian sufficient statistics [Liu, 2016a]. We assume the released data is of the same sample size as the original data \mathbf{x} in all the dips methods presented below. Among the above mentioned dips methods, only the modips method suggests releasing $m > 1$ sets surrogate synthetic data. However, releasing of multiple surrogate data sets mandates each set receives only $1/m$ of the total privacy budget. This suggested allocation of privacy budget prevents surpassing the prespecified overall privacy budget across all m released data sets per the sequential mechanism [McSherry and Talwar, 2007]. Even though the privacy budget is divided in this manner, generating the multiple datasets is a necessary step to account for the uncertainty of a synthesis model. We suggest the same principle should be applied to other dips approaches whenever the parameters in the predictive distribution for generating the synthetic data $\tilde{\mathbf{x}}$ is not fixed. In such cases, releasing multiple sets allows us to take into account the uncertainty of the parameters in the inferences based on the synthetic data. If the synthetic data are simulated from a single fixed distribution, then releasing multiple sets is not necessary. We will comment on which dips methods presented below are subject to multiple release.

3.1 np-dips

If np-dips is applied to categorical data, the statistics, \mathbf{s} , targeted for sanitization are proportions of the cells in the full cross-tabulation of the categorical data, \mathbf{x} . In the case of continuous data, the np-dips techniques generate differentially private smoothed density histograms, perturbed density histograms or release differentially private empirical distributions via the exponential mechanism.

3.1.1 Categorical data

The most straightforward approach in sanitizing categorical data \mathbf{x} is to add Laplace noises to the counts of the cross-tabulation of \mathbf{x} . Denote the multinomial distribution with K cells formed by the cross-tabulation of \mathbf{x} by $f(\mathbf{n}|\boldsymbol{\pi}) = \text{M}(n, \boldsymbol{\pi})$, where $n = (n_1, \dots, n_K)$. The Laplace sanitizer perturbs the original \mathbf{n} via $n_j^* = n_j + r_j$, where $r_j \sim \text{Lap}(0, \delta_s/\epsilon)$, and δ_s is the GS of releasing the multinomial data. δ_s can be set at 2 [Abowd and Vilhuber, 2008] or 1 [Dwork and Roth, 2013], depending on how $\delta(\mathbf{x}, \mathbf{x}') = 1$ is defined – change in one observation (thus $\delta_s = 2$) versus a removal of one observation (thus $\delta_s = 1$). For the remainder of the paper, we assume $\delta(\mathbf{x}, \mathbf{x}') = 1$ defined by the latter (removal of one observation). Abowd and Vilhuber [2008] proposed the Multinomial-Dirichlet (MD) synthesizer to generate differentially private synthetic data sets in the Bayesian framework. Assume the prior on $\boldsymbol{\pi}$ follows a Dirichlet distribution $f(\boldsymbol{\pi}) = \text{D}(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is a $(K \times 1)$ vector of prior sample sizes. The MD synthesizer sets the prior parameters at $\alpha_j^* = \frac{n}{\exp(\epsilon)-1}$ for $j = 1, \dots, K$, the minimum value that guarantees ϵ -differential privacy. $\boldsymbol{\pi}^*$ is first simulated from the posterior distribution $f(\boldsymbol{\pi}^*|\mathbf{x}) = \text{D}(\boldsymbol{\alpha}^* + \mathbf{n})$, and then the categorical synthetic data set is drawn from $f(\tilde{\mathbf{x}}|\boldsymbol{\pi}^*) = \text{M}(\mathbf{n}, \boldsymbol{\pi}^*)$.

The MD synthesizer motivates two approaches for synthesizing binary data in Charest [2010] and McClure and Reiter [2012], respectively. The MD synthesizer reduces to a Binomial-Beta (BB) synthesizer in Charest [2010]. That is, $f(p^*|\mathbf{x}) = \text{Beta}(\alpha_1 + n_1, \alpha_2 + n - n_1)$ and $f(\tilde{\mathbf{x}}|p^*) = \text{Binomial}(n, p^*)$, where $n_1 = \#\{x_i = 1\}$ in the binary data $\mathbf{x} = \{x_1, \dots, x_n\}$, $\alpha_1 = \alpha_2$ are set at $\frac{n}{\exp(\epsilon)-1}$ to satisfy ϵ -DP. McClure and Reiter [2012] synthesized data $\tilde{\mathbf{x}} = \{\tilde{x}_1, \dots, \tilde{x}_n\}$ by drawing from $f(\tilde{\mathbf{x}}|\mathbf{x}) = \text{Bernoulli}\left(\frac{n_1 + \alpha_\epsilon}{n + \alpha_\epsilon + \beta_\epsilon}\right)$, where $\alpha_\epsilon = \beta_\epsilon = (e^{\epsilon/n} - 1)^{-1}$, satisfying ϵ -DP. We refer to the latter approach as the DP-prior synthesizer. The DP-prior synthesizer simulates synthetic data from a distribution with fixed parameters given \mathbf{x} whereas the parameters in the BB or MD synthesizers are drawn from either a Beta or a

Dirichlet distribution. Since \mathbf{x} is drawn from a fixed distribution in the DP-prior method, there is no need to release multiple data sets. The BB synthesizer is a special case of the MD synthesizer, and a new p^* is simulated for each surrogate set of \mathbf{x} , so multiple release is suggested.

It should be noted that the prior information used in the MD synthesizer, the BB synthesizer, and the DP-prior methods increases with the data sample size n . Since the prior mean is 0.5, this implies the inferences based on the synthetic data will be biased if the true proportion is not 0.5, regardless how large n is. As a result, [Charest \[2010\]](#) proposed to model explicitly the data generation mechanism in a Bayesian framework from which inferences about the proportion can be obtained from a single set of synthetic set.

3.1.2 Numerical data

[Wasserman and Zhou \[2010\]](#) presented three different approaches of sampling from a differentially private empirical distribution to release synthetic data: perturbed histogram, smoothed histogram via the Laplace mechanism, and differentially private empirical distributions via the Exponential mechanism. Let h denote the bin width of a histogram, K represent the total number of bins, and $\{B_1, \dots, B_K\}$ be the bins of width h , $C_j = \sum_{i=1}^n I(x_i \in B_j)$ be the number of observations in B_j , and $\hat{p}_j = C_j/n$, where $j = 1, \dots, K$ and $I()$ is the indicator function (if $x_i \in B_j$, $I(x_i \in B_j) = 1$; 0 otherwise), a mean-squared consistent density histogram estimator is given by [\[Scott, 2015\]](#)

$$\hat{f}_K(x) = \sum_{j=1}^K K \hat{p}_j I(x \in B_j), \quad (3)$$

Sampling from a perturbed histogram is a direct application of the Laplace mechanism. The sanitized bin counts and proportions with ϵ -DP are given by $C_j^* = C_j + r_j$ and $\hat{p}_j^* = C_j^* / \sum_j C_j^*$, respectively, where $r_j \sim \text{Lap}(0, \delta_s / \epsilon)$ with $\delta_s = 1$. Since C_j^* preserves differential privacy, so does $(\hat{p}_1^*, \dots, \hat{p}_K^*)$

[Wasserman and Zhou, 2010]. Note that C_j^* can be negative if the regular Laplace mechanism is used.

A common used post-hoc processing approach is to replace negative C_j^* with 0 [Barak et al., 2007].

The density histogram estimator that satisfies ϵ -DP is thus given by

$$\hat{f}_K^*(x) = \sum_{j=1}^K K \hat{p}_j^* I(x \in B_j). \quad (4)$$

Since p_j^* is simulated from a distribution rather than fixed at a single value, releasing multiple sets of $\tilde{\mathbf{x}}$ is suggested, at different draws of p_j^* .

The smoothed histogram originally defined in Wasserman and Zhou [2010] that satisfies ϵ -DP applies to data bounded from 0 to 1. We extend the smoothed histogram to \mathbf{x} bounded by $[c_{i0}, c_{i1}]$, where $i = 1, \dots, p$ and p is the total number of attributes. The generalized smoothed histogram is given by

$$\hat{f}_K^*(x) = (1 - \lambda) \hat{f}_K(x) + \lambda \Delta, \quad (5)$$

where $\Delta = [(c_{11} - c_{10}) \times \dots \times (c_{p1} - c_{p0})]^{-1}$ and λ is a constant between 0 and 1

$$\lambda \geq \frac{K}{K + n(e^{\epsilon/n} - 1)} \quad (6)$$

to satisfy ϵ -DP. When $\epsilon \rightarrow 0$, $\lambda \rightarrow 1$ and the synthetic data are to be simulated from a uniform-like $\hat{f}_K^*(x)$ — too noisy to be of any use. When $\epsilon \rightarrow \infty$, $\lambda \rightarrow 0$ and $\hat{f}_K^*(x) \rightarrow \hat{f}_K(x)$, implying the synthetic data are to be simulated from the original density histogram $\hat{f}_K(x)$ with minimal privacy protection. Since $\hat{f}_K(x)$ is determined once the λ is specified in the smoothed histogram approach, it is not necessary to release multiple sets of $\tilde{\mathbf{x}}$, and a single synthetic set would be sufficient in the smoothed histogram approach.

To simulate synthetic data from the differentially private histograms given in Equations (4) and (5), both of which are piecewise uniform distributions, we can first draw a bin according to the relative

frequencies of the bins, then draw a point randomly from the bin obtained in the previous step.

In the approach of releasing differentially private empirical cumulative density functions (CDF) via the Exponential mechanism, synthetic data $\tilde{\mathbf{x}}$ is simulated from $h(\tilde{\mathbf{x}}|\mathbf{x})$ defined as

$$h(\tilde{\mathbf{x}}|\mathbf{x}) = \frac{g_{\mathbf{x}}(\tilde{\mathbf{x}})}{\int_{[c_{10}, c_{11}] \times \dots \times [c_{p0}, c_{p1}]} g_{\mathbf{x}}(\mathbf{z}) d\mathbf{z}}, \quad (7)$$

$$g_{\mathbf{x}}(\tilde{\mathbf{x}}) = \exp\left(-u(\hat{F}_{\mathbf{x}}, \hat{F}_{\tilde{\mathbf{x}}}) \frac{\epsilon}{2\delta_u}\right), \quad \delta_u = \sup_{\mathbf{x}, \mathbf{x}', \delta(\mathbf{x}, \mathbf{x}')=1} \sup_{\tilde{\mathbf{x}}} \left| u(\hat{F}_{\mathbf{x}}, \hat{F}_{\tilde{\mathbf{x}}}) - u(\hat{F}_{\mathbf{x}'}, \hat{F}_{\tilde{\mathbf{x}}}) \right|,$$

where $\hat{F}_{\mathbf{x}}$ is the original empirical CDF, $\hat{F}_{\tilde{\mathbf{x}}}$ is the empirical CDF's of the sanitized data, u is the utility function that denotes a distance measure between the two CDFs, and δ_u is the sensitivity of u . If the Kolmogorov-Smirnov distance is used on u , $\delta_u \leq n^{-1}$. However, releasing via the Exponential mechanism defined in Equation (7) does not seem to be a viable choice in practice. One reason for the difficulty lies in defining the outcome set of all possible candidate CDFs, the number of which increases rapidly with sample size n and the number of variables in \mathbf{x} , making releasing synthetic data from Equation (7) unrealistic for a large data set from the computational perspective. Strictly speaking, what is released from Equation (7) is an empirical CDF $\hat{F}_{\tilde{\mathbf{x}}}$ rather than $\tilde{\mathbf{x}}$ itself, so an extra step is needed to draw \tilde{x} from $\hat{F}_{\tilde{\mathbf{x}}}$. Given this, generation of multiple synthetic data becomes necessary to capture the uncertainty around $\hat{F}_{\tilde{\mathbf{x}}}$.

3.2 modips

The modips approach is based in a Bayesian modeling framework and releases m multiple sets of surrogate copies of the original data \mathbf{x} to account for the uncertainty of the synthesis model [Liu, 2016a]. An illustration of the steps of the modips algorithm is given in Figure 1. In summary, the modips sanitizes the Bayesian sufficient statistics \mathbf{s} from an appropriately constructed Bayesian model

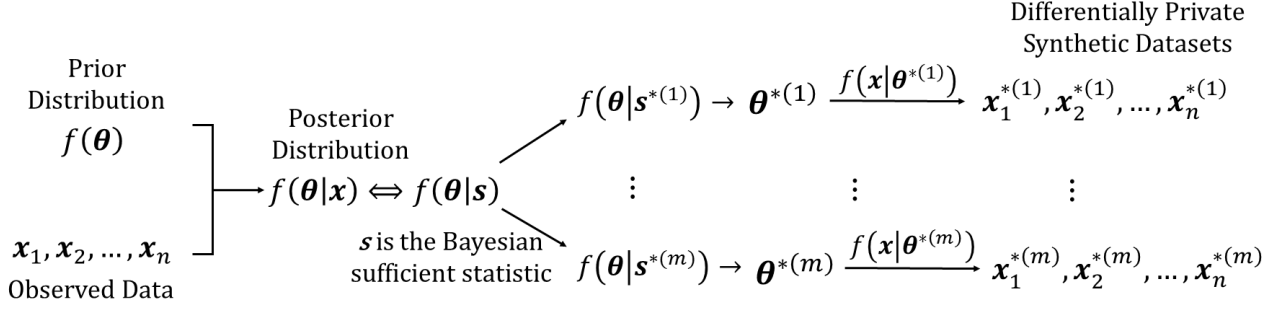


Figure 1: The modips algorithm.

$f(\boldsymbol{\theta}|\mathbf{s})$ with privacy budget ϵ/m , where m is the number of surrogate data sets. Denote the sanitized \mathbf{s} by \mathbf{s}^* . Synthetic data $\tilde{\mathbf{x}}$ is simulated given \mathbf{s}^* by first drawing $\boldsymbol{\theta}^*$ from the posterior distribution $f(\boldsymbol{\theta}|\mathbf{s}^*)$, and then simulating $\tilde{\mathbf{x}}$ from $f(\mathbf{x}|\boldsymbol{\theta}^*)$. The procedure is repeated m times to generate m surrogate data sets $\tilde{\mathbf{x}}$.

3.3 Inferences from synthetic data via dips

In the case of releasing multiple sets of synthetic data, such as using the MD synthesizer and the modips approach, inferential rules are needed to combine the inferences from the multiple set to yield an overall estimate. Suppose the parameter of interest is β . Denote the estimate of β in the j^{th} synthetic data by $\hat{\beta}_j$ and the associated standard error by v_j . The final inferences on β are obtained via

$$\bar{\beta} = m^{-1} \sum_{j=1}^m \hat{\beta}_j \quad (8)$$

$$T = m^{-1} B + W \quad (9)$$

$$(\bar{\beta} - \beta) T^{-1/2} \sim t_{\nu=(m-1)(1+mW/B)^2}, \quad (10)$$

where $B = (1 + m^{-1})(m - 1) \sum_{i=1}^m (\hat{\beta}_j - \bar{\beta})^2$ (between-set variability) and $W = m^{-1} \sum_{j=1}^m v_j^2$ (average per-set variability). The variance combination rule given in Equation (9) was proposed by Reiter [2003] for dealing with inferences in the context of partial synthesis without differential privacy. Releasing a full synthetic data set from a synthesis model build upon sample data \mathbf{x} model is the special case of partial synthesis with the synthesis portion equal to 100% [Liu, 2016a]. Therefore, the rule applies in the full synthesis for the traditional MS approach. Liu [2016a] proved that the combination rule in Equation (9) also applies in the modips approach. B in the modips approach comprises of two sources of variability – one from sanitizing \mathbf{s} and another from drawing the model parameters from their posterior distribution given sanitized \mathbf{s}^* . Due to the extra sanitization step of \mathbf{s} in the modips approach as compared to the traditional MS approach, B in the former will be larger than in the latter, leading to less precise estimate on β , a price paid for differential privacy guarantee. The variance combination rule given in Equations (8) to (10) also applies in the MD synthesizer and other dips approaches that rely on multiple set releases to account for synthesis model uncertainty. In fact, the MD synthesizer can be regarded as a modips approach for categorical data.

4 Simulation Studies

We assess the utility and inferential properties of the sanitized data via the dips approaches presented in Section 3 in three simulation studies. The first and second simulation studies focus on categorical data and continuous data generated from binomial and normal distributions, respectively; the third simulation involves a mixture data set of both categorical and continuous variables. The results are benchmarked against the inferences based on the original data before sanitization and the traditional MS technique without differential privacy. In all synthesis approaches examined, the sample size of the released synthetic data is the same as the original data.

4.1 Simulation study 1: categorical data

The following methods are compared in the categorical data simulation study: the modips synthesizer, the Laplace sanitizer, the DP-prior synthesizer, and the MD synthesizer (reduced to the Binomial-Beta synthesizer in the binomial case). Data was simulated from a Bernoulli distribution $f(x_i) = \text{Bern}(\pi)$ for $i = 1, \dots, n$. We examined 9 simulation scenarios for $n = \{40, 100, 1000\}$ and $\pi = \{0.10, 0.25, 0.50\}$. In all the dips approaches, we examined the effect of privacy budget on the statistical inferences based on the synthetic data by varying $\ln(\epsilon)$ at $\{-10, -9, -8, \dots, 8, 9, 10\}$.

In the Bayesian framework, if $\text{Beta}(\alpha, \beta)$ is employed as the prior on π , then the posterior distribution of π given \mathbf{x} is $f(\pi|\mathbf{x}) = \text{Beta}(\alpha + n_1, \beta + n - n_1)$, where $n_1 = \#\{x_i = 1\}$, and the posterior predictive distribution is $f(\tilde{x}_i|\mathbf{x}) = \int f(x_i|\pi)f(\pi|\mathbf{x})d\pi$. In this simulation, we set $\alpha = \beta = 1/3$ in the prior distribution of π [Kerman, 2011]. For the traditional MS, we sampled π from $f(\pi|n_1) = \text{Beta}(\alpha + n_1, \beta + n - n_1)$, and \tilde{x}_i from $f(\tilde{x}_i|\pi) = \text{Bern}(\pi)$ for $i = 1, \dots, n$. The cycle was repeated 10 times (drawing π and then drawing $\tilde{\mathbf{x}}$) to obtain 10 sets of synthetic binary data. In the modips approach, we first located the Bayesian sufficient statistics \mathbf{s} associated with $f(\pi|\mathbf{x}) = \text{Beta}(\alpha + n_1, \beta + n - n_1)$, which, in this case, was n_1 with $\delta_{\mathbf{s}} = 1$; the Laplace mechanism was then employed to add noise, e , to obtain $n_1^* = n_1 + e$, where $e \sim \text{Lap}(0, \epsilon^{-1})$. Given the sanitized n_1^* , the modips sampled π^* from $f(\pi^*|n_1^*) = \text{Beta}(\alpha + n_1^*, \beta + n - n_1^*)$, and \tilde{x}_i from $f(\tilde{x}_i|\pi^*) = \text{Bern}(\pi^*)$ for $i = 1, \dots, n$ to generate a set of synthetic data. The cycle was repeated 10 times (from sanitizing n_1 to drawing $\tilde{\mathbf{x}}$) to obtain 10 sets of synthetic binary data. Since the e was drawn from Laplace distribution with a support from $(-\infty, \infty)$, the sanitized n_1^* could be out of bounds of $[0, n]$. There are two data-independent post-processing ways: First, to truncate the distribution \mathbf{e} at the boundaries of 0 and n ; second, to set

values < 0 at 0 and $> n$ at n , resulting in a piecewise distribution. We refer to the former approach by “truncation” and the second by “thresholding”. $p_1^* = n_1^*/n$ is asymptotically unbiased for $p_1 = n_1/n$. When n is small, the thresholding approach is less biased than the truncation approach [Liu, 2016a]. We examined both approaches in this simulation.

In the Laplace sanitizer, $n_1^* = n_1 + e$, where $e \sim \text{Lap}(0, \epsilon^{-1})$, and a single data set with $n_1^* = \{\#x_i = 1\}$ was released. Similar to the modips approach, n_1^* can be out of bounds, the thresholding approach was applied when this happened. Both the DP-prior synthesizer and the MD synthesizer simulate the data from $\tilde{\mathbf{x}} \sim \text{Bern}(p^*)$; however, p^* was fixed at $\frac{n_1 + \alpha^*}{n + \alpha^* + \beta^*}$ with $\alpha^* = \beta^* = (e^{\epsilon/n} - 1)^{-1}$ for the DP-prior synthesizer, and was drawn from $f(p^* | \alpha^*, \beta^*) = \text{Beta}(\alpha^* + n_1, \beta^* + n - n_1)$ for the MD synthesizer with $\alpha_\epsilon = \beta_\epsilon = (e^{\epsilon/n} - 1)^{-1}$.

A single synthetic data set was generated from the DP-prior synthesizer and 10 sets of synthetic data were generated from the MD synthesizer. For the traditional MS and the dips approaches that released multiple synthetic data sets, each synthetic data set received 1/10 of the total privacy budget per the sequential mechanism. To obtain inference on π from the released data set, each of the 10 sets was analyzed separately to obtain a point estimate of π , which was the sample proportion, \hat{p}_j , where the within-set variance of \hat{p}_j , v_i , was calculated as $\hat{p}_j(1 - \hat{p}_j)n^{-1}$ in each set for $j = 1, \dots, 10$, and Equations (8) to (10) were applied to obtain a final estimate of \hat{p} and the 95% confidence interval (CI).

Figures 2 to 3 depict the results on the bias, root mean squared error (RMSE), and coverage probability (CP) of the 95% CI in the inferences for π based on the synthetic data from each dips approach as well as the MS approach and the original data. In all dips approaches, there was some noticeable bias, large RMSE, and some undercoverage at small ϵ ($\epsilon < 1$). The inferences improved as ϵ increases (more privacy budget and less perturbation), and eventually approached the original or the MS results (except

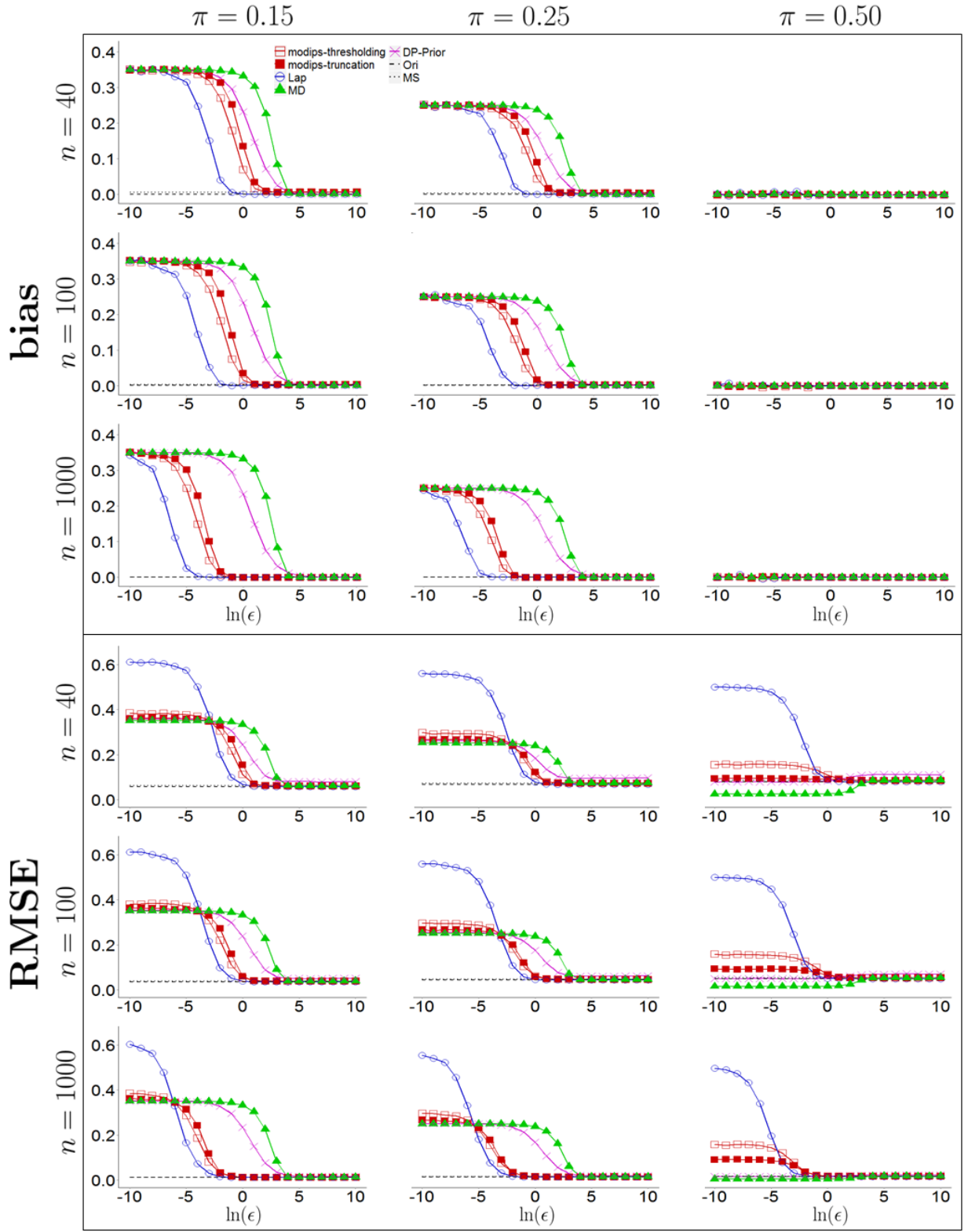


Figure 2: Bias and RMSE of π . *modips* represents the model-based differentially private synthesis, *Lap* represents the Laplace synthesizer, *MD* represents the MD synthesizer, *DP-prior* represents DP prior synthesizer, *Ori* is the original results without any perturbation, and *MS* is the traditional multiple synthesis method without DP.

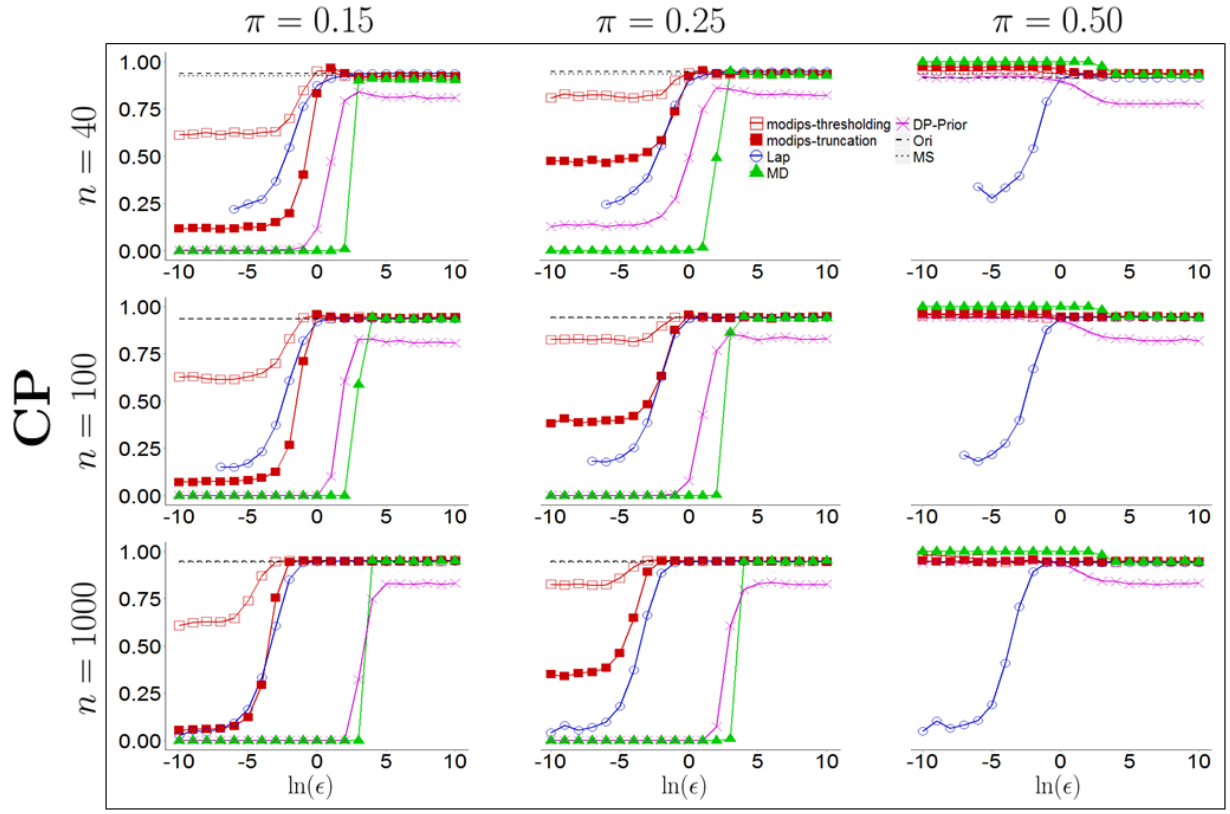


Figure 3: Coverage Probability (CP) of π . *modips* represents the model-based differentially private synthesis, *Lap* represents the Laplace synthesizer, *MD* represents the MD synthesizer, *DP-prior* represents DP-prior synthesizer, *Ori* is the original data results, and *MS* is the traditional multiple synthesis method without DP (CP in *Lap* for $\ln(\epsilon)$ was not plotted due to the small number of usable simulated data sets (Table 1 in Supplemental Materials))

for the DP-prior synthesizer). Specifically, the inferential results from the *modips* and the Laplace sanitizer approached those in the traditional MS approach, and the MD synthesizer approached the original data's results. The inferences in the *modips* approach and the Lap synthesizer improved as n increased since the synthesizers in both were the same regardless of n ($\delta_s = 1$, independent of n). As a result, the amount of perturbation becomes less relative to the size of the data as n increased. In the case of the MD synthesizer and the DP-prior synthesizer, since the prior information increased with n ($n/(e^\epsilon - 1)$ and $1/(e^{\epsilon/n} - 1)$, respectively) and the prior mean of π was 0.5. This caused the bias remained large when $\pi \neq 0.5$ regardless of n . Also, the RMSE was smaller than the original and the CI over-covered when $\pi = 0.5$. Among the three π values examined, the inferences were the best when $\pi = 0.5$ in all dips approach. In the case of the MD synthesizer and the DP-prior synthesizer, it

can be explained by the consistent between the prior information (ever increasing with n) and data; In the case of the Laplace synthesizer and the modips approach, it can be explained by the fact that $p = 0.5$ is the center point of the support of a proportion, either truncating at 0 or 1 or thresholding at 0 and 1 does not skew the distribution of p .

In terms of bias, when $\pi = 0.15$ and $\pi = 0.25$, the MD synthesizer was the worst and the Laplace synthesizer was the best. As expected, modips-truncation was worse than modips-thresholding, but both were better than the DP-prior approach. The differences in bias among the dips approach were minimal when $\pi = 0.5$ except for the large fluctuation in the Laplace synthesizer at small ϵ (Figure 1 in Supplemental Materials). In terms of RMSE, the relative magnitudes between the Laplace synthesizer and the MD synthesizer switched at a certain value of ϵ , depending on the values of π and n . For example, when $\pi = 0.15$ and $n = 40$, the RMSE from the Laplace synthesizer was the largest and the RMSE from the MD synthesizer was the smallest when $\epsilon < e^{-5}$; when $\epsilon > e^{-5}$, the RMSE from the Laplace synthesizer was the smallest and the RMSE from the MD synthesizer was the largest. The RMSE from the modips and the DP-prior approach were also in between the former two approaches, with the modips-truncation approach having slightly larger RMSE than the modips-thresholding approach, and both better than the DP-prior approach. In terms of CP, the modips-thresholding approach had the most proper coverage (closest to the nominal level 0.95) regardless of n , π , and ϵ . When $\pi = 0.10$ and 0.25 , for the Laplace synthesizer, the MD synthesizer, and the DP-prior synthesizer there was better coverage at small ϵ with smaller n than with larger n values. Every modips approaches suffered from some degree of undercoverage at small ϵ due the relatively large bias at those ϵ values; the undercoverage improved when π approaches 0.50 and n becomes large. While the CP in all the other dips techniques approached the nominal level of 95%, the CP of the DP-prior

did not. This is due to the fact that the synthesis model is fixed for a given ϵ and \mathbf{x} , and it does not promote the uncertainty of the synthesis model in an appropriate manner. When $\pi = 0.10$ and 0.25 , in all dips approaches except for the modips thresholding synthesizer, there was better coverage at smaller n when ϵ was small. This result was probably due to the fact that the bias and RMSE did not improve until $\ln(\epsilon) \approx 1$ while the CI continued to shrink as n increased, leading to worse coverage. Figure 2 in Supplemental Materials also presents the average CI width for each approach. Altogether, the modips method performed the best among all the dips approaches, due to its consistent performance across all ϵ values and improved inferences with increased n .

4.2 Simulation study 2: continuous data

The following methods are compared in the continuous data simulation study: the modips synthesizer, the np-dips synthesizer via the perturbed histogram and the smoothed histogram approaches. Data was simulated from $N(\mu, \sigma^2)$, for $i = 1, \dots, n$. Since DP was defined in the context of bounded data, we manually truncated the simulated data at $[a = \mu - 3\sigma, b = \mu + 4\sigma]$. Since there was minimal probability mass (0.0013) outside the -3σ and 4σ range, the normal assumption was hardly affected with the truncation. We examined 9 simulation scenarios for $n = \{20, 100, 1000\}$ and $\sigma^2 = \{1, 4, 9\}$, with 5000 repetitions per scenario. Without loss of generality, μ was set to 0 in all scenarios. Furthermore, we varied the privacy budget $\ln(\epsilon)$ at $\{-10, -9, -8, \dots, 8, 9, 10\}$.

For the traditional MS method, we assumed prior $f(\mu, \sigma^2) \propto \sigma^{-2}$. The posterior distributions were thus $f(\sigma^2|\mathbf{x}) = \text{Inv-Gamma}\left(\frac{n-1}{2}, \frac{(n-1)S^2}{2}\right)$ and $f(\mu|\mathbf{x}, \sigma^2) = N(\bar{x}, n^{-1}\sigma^2)$, where \bar{x} and S^2 were the sample mean and variance, respectively. The posterior predictive distribution was $f(\tilde{x}_i|\mathbf{x}) = \int f(\tilde{x}_i|\mu, \sigma^2)f(\mu|\sigma^2, \mathbf{x})f(\sigma^2|\mathbf{x})d\mu d\sigma^2$. We generated one set of surrogate data by first drawing σ^2 and

μ from their posterior distribution, and then drawing $\tilde{\mathbf{x}}$ from the normal distribution given the drawn μ and σ^2 . $m = 10$ sets of synthetic data were simulated to release.

The Bayesian sufficient statistics from the posterior distributions (μ, σ^2) was $\mathbf{s} = (\bar{x}, S^2)$. The modips procedure started with sanitizing \mathbf{s} via the Laplace mechanism to obtain \mathbf{s}^* . Specifically, the GS was $(b - a)n^{-1}$ for \bar{x} and $(b - a)^2n^{-1}$ for S^2 , where $(b - a) = 7\sigma$ [Liu, 2016a]. Note that \bar{x} ranged $[a, b]$, and S^2 ranged $\left[0, \frac{(b-a)^2}{4} \frac{n}{n-1}\right]$. If a sanitized \mathbf{s}^* was outside the range, it was replaced by the boundary values (modips-thresholding). Given the sanitized $\mathbf{s}^* = \{\bar{x}^*, S^{2*}\}$, the modips technique then drew σ^{2*} from Inv-Gamma $\left(\frac{n-1}{2}, \frac{(n-1)S^{2*}}{2}\right)$ and μ^* from $N(\bar{x}^*, n^{-1}\sigma^{2*})$. Finally, \tilde{x}_i was simulated from $N(\mu^*, \sigma^{2*})$ for $i = 1, \dots, n$ to generate one set of surrogate data. The whole procedure was repeated $m = 10$ times to generate 10 surrogate data sets. To keep the total privacy budget at ϵ , $1/10$ of the total budget was spent when synthesizing one of the 10 surrogate data sets. In addition, since there were two statistics, (\bar{x}, S^2) , to sanitize over the same set of data, the rate parameter in the Laplace distribution (from which the noise terms added to \bar{x} and S^2 were drawn) needed to be appropriately configured. We applied two approaches in this simulation. In the “non-split” approach, the default Laplace mechanism, ϵ -DP can be preserved by perturbing each query from a query sequence q_1, \dots, q_r , where r is the total number of queries by adding a noise term from $\text{Lap}(\delta_{\mathbf{s}}\epsilon^{-1})$, and where $\delta_{\mathbf{s}} = \sum_{k=1}^r \delta_{q_k}$. In the “split” approach, motivated by the sequential composition principle, the total privacy budget ϵ among the r queries is split to r equal portions, and query k is perturbed with an added noise term from the $\text{Lap}(\delta_{q_k}(\epsilon/r)^{-1})$. In the context in this simulation, $\delta_{\mathbf{s}} = \delta_{\bar{x}} + \delta_{S^2} = \frac{b-a+(b-a)^2}{n}$ and $\epsilon/10$ (ϵ is total privacy budget) in the “non-split” approach. In the modips-split approach, we allocated half of the privacy budget to sanitize \bar{x} and the other half to sanitize S^2 , and then applied the Laplace mechanism to obtain \bar{x}^* and S^{2*} , separately.

In deciding the number of bins for the histograms in the perturbed and smoothed histogram approaches, we applied the Scott's Rule based on the comparison results among the Sturges' Rule, the Scott's Rule, and the F-D rule [Scott, 2015]. Specifically, the bin width was $\hat{h} = 3.5Sn^{-1/3}$, where S was the sample standard deviation of a variable in \mathbf{x} and n was the sample size. The median number of bins was 7, 10, and 21 for $n = 20, 100$, and 1000 , respectively, across all simulations (Table 2 in Supplemental Materials). The number of bins remained fairly constant across σ^2 since the range of \mathbf{x} was proportional to standard deviation (7σ). Each bin count was perturbed via the Laplace mechanism with $\delta_s = 1$ to obtain the perturbed density histogram from Equation (4). 10 sets of synthetic data were then simulated from the perturbed density histogram; one for each differentially private $\hat{\mathbf{p}}^*$ via the Laplace mechanism. For the smoothed histogram, we first calculated λ for a given ϵ using Equation (6), and then constructed the smoothed histogram by applying Equation (5), from which a single set of synthetic data was generated and released.

When obtaining inference on μ from the multiple released data sets (the modips approach, the perturbed histogram, and the MS approach), each of the 10 synthetic sets was analyzed separately to obtain a point estimate of μ , which was \bar{x} ; v_j , the per-set variance of \bar{x} , was estimated S_j^2/n , where S_j^2 was the sample variance in synthetic data set j for $j = 1, \dots, 10$. Equations (8) to (10) were then applied to obtain a final estimate of μ and the 95% CI.

Figures 4 and 5 depict the bias, RMSE, and CP in each simulation scenario of μ for the modips-split, modips-nonsplit, perturbed histogram, and smoothed histogram approaches. The modips-split and modips-nonsplit methods performed the best among all the dips approaches, with the former slightly better. In all dips approaches, there was some noticeable bias and large RMSE at small ϵ . Both bias and RMSE improved as ϵ increases (more privacy budget and less perturbation), and eventually

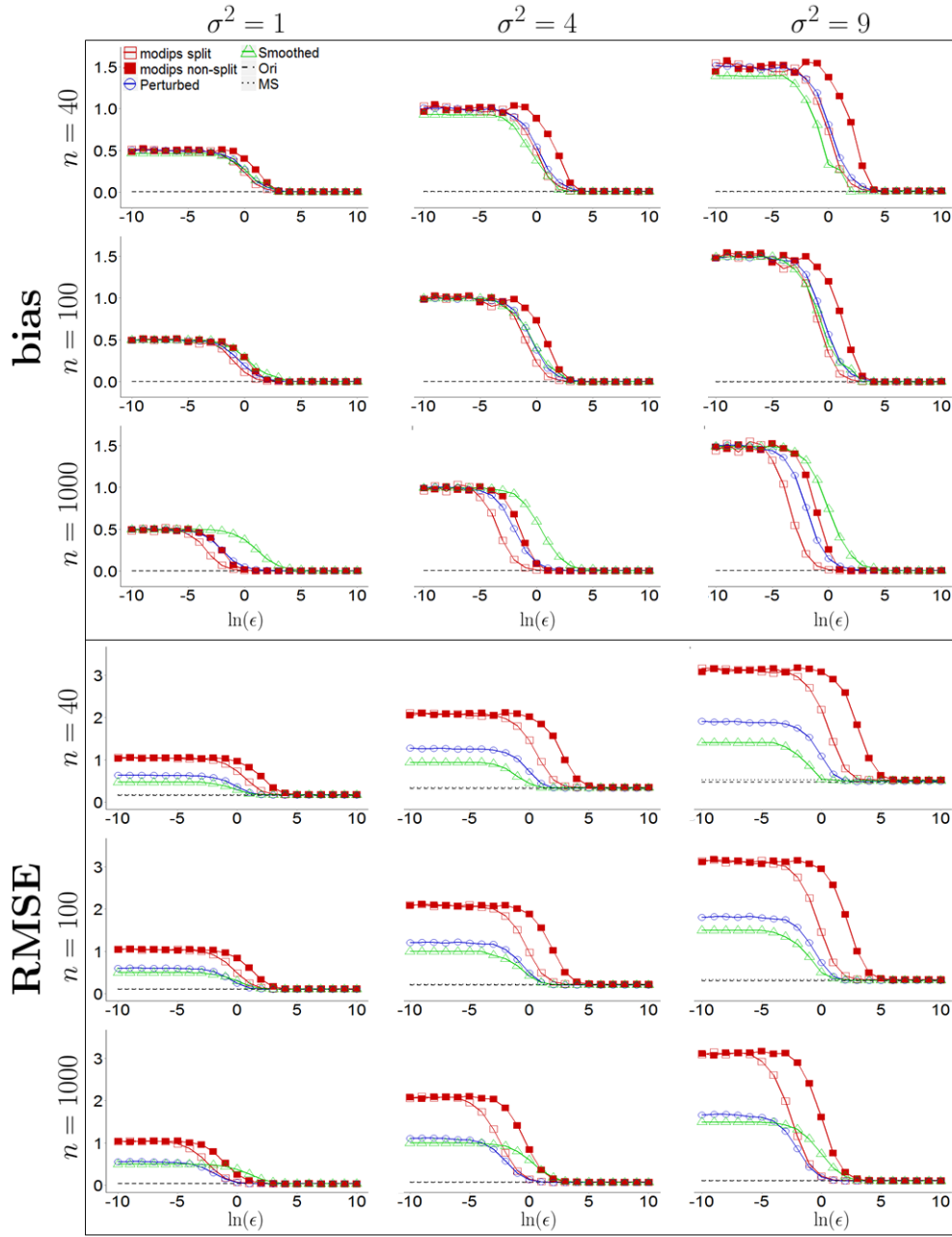


Figure 4: Bias and RMSE of μ . *modips-split* and *modips-nonsplit* (with thresholding) represent the model-based differentially private synthesis with ϵ split and not split (respectively) between μ and σ^2 , *perturbed* represents the perturbed histogram method, *smoothed* represents the smoothed histogram method, *Ori* is the original results without any perturbation, and *MS* is the traditional multiple synthesis method without DP.

approached the original or the MS results. Since the global sensitivity of \bar{x} and S^2 became smaller as n increased, the bias, RMSE, and CP approached the desired values at a quicker rate for larger n in the *modips-split*, *modips-nonsplit*, and *perturbed* histogram methods. However, the smoothed histogram performed worse for larger values of n . Recall that the smoothed histogram defined in Equation (5) generates synthetic data from a uniform-like $\hat{f}_K^*(x)$ when $\lambda \rightarrow 1$. How, as $n \rightarrow \infty$, $\lambda \rightarrow 1$ from

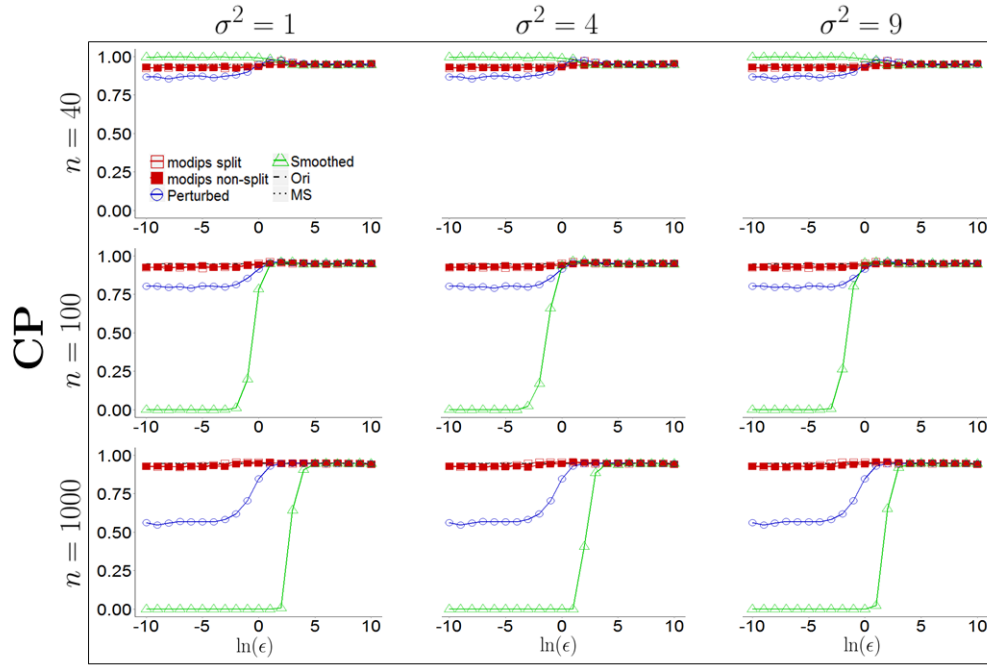


Figure 5: Coverage Probability (CP) of μ . *modips-split* and *modips-nonsplit* (with thresholding) represent the model-based differentially private synthesis with ϵ split and not split (respectively) between μ and σ^2 , *perturbed* represents the perturbed histogram method, *smoothed* represents the smoothed histogram method, *Ori* is the original results without any perturbation, and *MS* is the traditional multiple synthesis method without DP.

Equation (6), resulting in a poor performance for larger values of n .

The bias was positive in all examined cases, implying μ was overestimated based on the synthetic data. This is because the bounds of data \mathbf{x} and $\bar{x} - [\mu - 3\sigma, \mu + 4\sigma]$ – were not symmetric about true value of μ . When synthetic \mathbf{x} and μ were out of bound, they were set at the boundary values. Since the left bound $\mu - 3\sigma$ is close to μ , the amount of values $< \mu - 3\sigma$ set at $\mu - 3\sigma$ was larger than values $> \mu + 4\sigma$ set at $\mu + 4\sigma$, resulting in overestimation. There were also some undercoverage at smaller values of ϵ in the two histogram based approaches for large n , but the two *modips* approaches had close-to-nominal level. The distinction between *modips-split* and *modips-nonsplit* becomes more obvious when σ^2 increased and n increased. The bias from the perturbed histogram and the smoothed histogram methods performed similarly at smaller values of ϵ , but the bias for the perturbed histogram approached 0 quicker as ϵ increased than the smoothed histogram in most scenarios. The RMSE in the histogram-based approaches were smaller than that from the *modips*, but it changed very little with increased n for smaller values of ϵ , so did the bias. For those values of ϵ , the widths of the

CIs kept shrinking with n , resulting in undercoverage. Figure 3 in Supplemental Materials presents the average CI width for each approach. As a comparison, we also examined the case when the data bounds were $[\mu - 4\sigma, \mu + 4\sigma]$, which was symmetric about \bar{x} . The results are presented in Figures 4 to 7 in Supplemental Materials. As expected, there were minimal biases on μ in all the dips approaches (there was more fluctuation around 0 in the modips approach). The RMSE was smaller in the histogram-based approaches while still offering coverage over or at the nominal level. In other words, the histogram-based approaches delivered more precise estimates than the modips approach in the inferences of μ . On the other hand, the histogram-based approaches did not perform as well as the modips approaches in the inferences of σ^2 – both bias and RMSE were large and there was severe undercoverage when ϵ was small.

We present the results on σ^2 in Supplemental Materials (Figures 8 and 9). The results were consistent with those on μ : the modips methods performed better than the two histogram-based approaches. In addition, the relative bias and RMSE in the former was much smaller than the latter across all σ^2 values and n values examined at smaller ϵ . The CP in the modips based approaches was also much better than those in the histogram-based approaches. Between the two modips approaches, the non-split approaches performed slightly better than the split approach, which was the opposite case of μ . The reason being that in the non-split approach, every statistic shared an aggregate δ_s , the sum of the individual δ_{s_k} . Statistics with $\delta_{s_k} < r^{-1}\delta_s$, the average of the individual GS, received more noise if they were sanitized separately via the split approach. In the context of simulation, $\delta_{\bar{x}} < \delta_S$. This means the split approach benefited from the inferences of μ while the non-split approach benefited from the inferences of σ^2 . Between the two histogram-based approaches, the perturbed approach was noticeably better than the smoothed case.

4.3 Simulation study 3: mixture model data

In this simulation study, we explored a more complicated data scenario where the data is a mixture of categorical and continuous variables. The modips synthesizer (split, thresholding) and the np-dips synthesizer via the perturbed histogram were compared. Let $\mathbf{x} = (\mathbf{w}, \mathbf{z})$, where \mathbf{w} denotes the categorical variables and \mathbf{z} denotes the continuous variables. Let n_k denote the count in each cell formed by the full cross-tabulation of \mathbf{w} , where $k = 1, \dots, K$, and K is the total number of cells. In this simulation, \mathbf{w} contained three categorical variables (w_1, w_2, w_3) with 2, 3, and 4 levels, respectively, yielding a total of $K = 24$ cells from the cross-tabulation of \mathbf{w} . The counts $\mathbf{n} = \{n_k\}$ in the 24 cells were simulated from a Multinomial distribution with parameter $\boldsymbol{\pi} = \{\pi_k\}$. The $\boldsymbol{\pi}$ values ranged from 0.007 to 0.076 such that $\sum_{k=1}^{24} \pi_k = 1$. $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)$ was simulated from the bivariate normal distribution $f(\mathbf{z}_{ki}) = N(\boldsymbol{\mu}_k, \Sigma)$ for $i = 1, \dots, n_k$, where $\boldsymbol{\mu}_k = \{\mu_{k1}, \mu_{k2}\}$ was the mean of \mathbf{z} in cell k , and Σ was the covariance matrix that was assumed to be the same across all 24 cells. The summary of the parameter values of $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\pi}$ across the 24 cells are provided in Table 3 in the Supplemental Materials. We set $n = 1000$, the variances of \mathbf{z}_{k1} and \mathbf{z}_{k2} , $\sigma_1^2 = \sigma_2^2 = 1$, and the correlation between \mathbf{z}_{k1} and \mathbf{z}_{k2} , $\rho = 0.50$ with 5000 repetitions. \mathbf{z}_{kj} in cell k (where $j = 1, 2$) was truncation at $[a_{kj} = \mu_{kj} - 4\sigma_j, b_{kj} = \mu_{kj} + 4\sigma_j]$ to generate bounded data so that the dips approaches could be applied. We also examined a range of privacy budget $\ln(\epsilon)$ at $\{-6, \dots, 8, 9, 10\}$.

For the traditional multiple synthesis approach, we imposed a Dirichlet prior on $\boldsymbol{\pi}$, $f(\boldsymbol{\pi}) = D(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = \{\alpha_1, \dots, \alpha_K\} = 1/2$, and its posterior distribution was $f(\boldsymbol{\pi}|\mathbf{w}) = D(\boldsymbol{\alpha}')$, where $\boldsymbol{\alpha}' = \boldsymbol{\alpha} + \mathbf{n}$ and $\mathbf{n} = \{n_1, \dots, n_K\}$. We applied noninformative priors $f(\boldsymbol{\mu}, \Sigma) \propto |\Sigma|^{-1}$ to $\boldsymbol{\mu}$ and Σ , and the posterior distributions were $f(\Sigma|\mathbf{z}, \mathbf{w}) = \text{Inv-Wishart}(n - K, \hat{\Sigma})$ and $f(\boldsymbol{\mu}_k|\Sigma, \mathbf{z}, \tilde{\mathbf{w}}) = N(\bar{\mathbf{z}}_k, \tilde{n}_k^{-1}\Sigma)$,

where $\hat{\Sigma} = \frac{1}{n} \sum_{k=1}^K \sum_{i=1}^{n_k} (\mathbf{z}_{ik} - \bar{\mathbf{z}}_k)(\mathbf{z}_{ik} - \bar{\mathbf{z}}_k)^T$, and $\bar{\mathbf{z}}_k$ was the sample mean of \mathbf{z} in cell k . 10 sets of synthetic data were simulated from the posterior predictive distribution $f(\tilde{\mathbf{z}}_i, \tilde{\mathbf{w}}_i | \mathbf{z}, \mathbf{w}) = \int f(\tilde{\mathbf{z}}_i | \boldsymbol{\mu}, \Sigma, \tilde{\mathbf{w}}_i) f(\boldsymbol{\mu} | \Sigma, \mathbf{z}, \mathbf{w}) f(\Sigma | \mathbf{z}, \mathbf{w}) f(\tilde{\mathbf{w}} | \boldsymbol{\pi}) f(\boldsymbol{\pi} | \mathbf{w}) d\boldsymbol{\pi} d\boldsymbol{\mu}_k d\Sigma$ for $i = 1, \dots, \tilde{n}_k$ and $k = 1, \dots, K$. First, we drew $\boldsymbol{\pi}$ from $f(\boldsymbol{\pi} | \mathbf{w}) = D(\boldsymbol{\alpha} + \mathbf{n})$, and $\tilde{\mathbf{w}}$ from $f(\tilde{\mathbf{w}} | \boldsymbol{\pi}) = \text{Multinomial}(n, \boldsymbol{\pi})$; next, we drew Σ from $f(\Sigma | \mathbf{z}, \mathbf{w}) = \text{Inv-Wishart}(n - K, \hat{\Sigma})$, $\boldsymbol{\mu}_k$ from $f(\boldsymbol{\mu}_k | \Sigma, \mathbf{z}, \mathbf{w}) = N(\bar{\mathbf{z}}_k, n_k^{-1} \Sigma)$, for $k = 1, \dots, 24$; finally, $\tilde{\mathbf{z}}_i$ was simulated from $f(\mathbf{z}_i | \tilde{\mathbf{w}}_i, \Sigma) = N(\boldsymbol{\mu}_{\tilde{k}_i}, \Sigma)$ for $i = 1, \dots, n$ to generate one set of surrogate data, where \tilde{k}_i indicates the cell which the simulated case i belongs to given the simulated \mathbf{w}_i .

The Bayesian sufficient statistics from the above Bayesian model was $\mathbf{s} = (\mathbf{n}, \hat{\Sigma}, \bar{\mathbf{z}})$, where $\bar{\mathbf{z}}$ was the cell sample means of \mathbf{z}_1 and \mathbf{z}_2 . The modips procedure started with sanitizing \mathbf{s} via the differentially private Laplace mechanism to obtain $\mathbf{s}^* = (\mathbf{n}^*, \hat{\Sigma}^*, \bar{\mathbf{z}}^*)$. The GS was 1 for \mathbf{n} , and $(b_{kj} - a_{kj})n^{-1}$ for \bar{z}_{kj} , where $b_{kj} - a_{kj} = 8\sigma$ for all $k = 1, \dots, 24$ and $j = 1, 2$; the GS for each entry in $\hat{\Sigma}$ was $(8\sigma)^2 n^{-1}$ [Liu, 2016a]. Given \mathbf{s}^* , the modips method first drew $\boldsymbol{\pi}^*$ from $f(\boldsymbol{\pi}^* | \mathbf{n}^*) = D(\boldsymbol{\alpha} + \mathbf{n}^*)$, $\tilde{\mathbf{w}}^*$ from $f(\tilde{\mathbf{w}} | \boldsymbol{\pi}^*) = \text{Multinom}(n, \boldsymbol{\pi}^*)$, Σ^* from $f(\Sigma^* | \hat{\Sigma}^*) = \text{Inv-Wishart}(n - K, \hat{\Sigma}^*)$, $\boldsymbol{\mu}_k^*$ from $f(\boldsymbol{\mu}_k^* | \Sigma^*, \bar{\mathbf{z}}^*, \mathbf{w}) = N(\bar{\mathbf{z}}_k^*, n_k^{-1} \Sigma^*)$; and then $\tilde{\mathbf{z}}_i$ was simulated from $f(\mathbf{z}_i | \boldsymbol{\mu}_{\tilde{k}_i}^*, \Sigma^*) = N(\boldsymbol{\mu}_{\tilde{k}_i}^*, \Sigma^*)$ for $i = 1, \dots, n$ to generate one set of surrogate data, where \tilde{k}_i indicates the cell which the simulated case i belongs to given the simulated \mathbf{w}_i . The whole procedure was repeated 10 times to generate 10 surrogate data sets. Each surrogate data set received 1/10 of the total privacy budget. Since \mathbf{s} contained 6 components: \mathbf{n} , $\bar{\mathbf{z}} = (\bar{\mathbf{z}}_1, \bar{\mathbf{z}}_2)$, two variance terms and one covariance from Σ , each of the 6 components received 1/6 of the privacy budget allocated to each synthetic data set in the modips-split approach.

In the np-dips approach, the Laplace sanitizer was first applied to sanitize the counts \mathbf{n} formed by the categorical \mathbf{w} , and the perturbed histogram method was then used to sanitize the continuous \mathbf{z} . Note that since \mathbf{z} was 2-dimensional, each histogram bin was a square rather than an interval. The Scott's

Rule suggest the medians of the number of bins ranged from 16 to 49 across the 24 cells (Supplemental Materials Table 4). The process was repeated 10 times to create 10 perturbed histograms, from which 10 sets of synthetic data were generated. As in the case of modips, the total budget was first split among the 10 data sets, and then was split in half again when generating each synthetic data given there were two queries (cell counts formed by \mathbf{w} , and the cell histograms formed by \mathbf{z}).

We examined the inferences on $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma$, and the marginal probabilities of \mathbf{w} based on the synthetic data set. We denote the marginal probabilities by $\boldsymbol{\Pi} = (\Pr(w_1 = 1), \Pr(w_2 = 1), \Pr(w_2 = 2), \Pr(w_3 = 1), \Pr(w_3 = 2), \Pr(w_3 = 3))$. $\boldsymbol{\Pi}$ was estimated by the sample marginal probabilities $\hat{\mathbf{P}}$; $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ were estimated by the sample cell means $\bar{\mathbf{z}}_1$ and $\bar{\mathbf{z}}_2$; and Σ was estimated by the pooled variance-covariance $\hat{\Sigma}$. The corresponding per-set variance for the point estimates were estimated by $\hat{\mathbf{P}}(1-\hat{\mathbf{P}})\mathbf{n}^{-1}, S_1^2 n_k^{-1}, S_2^2 n_k^{-1}, (S_1^2)^2(2(n-1)^{-1} + \kappa n^{-1}), (S_2^2)^2(2(n-1)^{-1} + \kappa n^{-1})$, and $(1-r^2)(n-2)^{-1}$, where S_1^2, S_2^2 were the diagonal elements of $\hat{\Sigma}$, κ was the excess kurtosis, and r was the correlation derived from $\hat{\Sigma}$. When obtaining inferences from the modips approach, the MS approach, and the perturbed histogram approach, each of the 10 sets was analyzed separately to obtain point estimates for $\boldsymbol{\Pi}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma$, and the corresponding within-set variance. Equations (8) to (10) were then applied to obtain the final estimates of the parameters and the 95% CIs.

Figure 6 shows the results on the bias, RMSE, and CP of the 24 cell means of \mathbf{z}_1 ($\boldsymbol{\mu}_1$) and the 6 marginal probabilities ($\boldsymbol{\Pi}$). The results on $\boldsymbol{\mu}_2$ and Σ are provided in Figures 10 and 11 in the Supplemental Materials. The inferences on $\boldsymbol{\mu}_1$ are presented in the format of a box plot over the 24 cells means. Overall, the modips technique performed better than the perturbed histogram in terms of $\boldsymbol{\Pi}$, where the bias and RMSE in the former were much smaller, and the CP was at or above the nominal level 95% in the modips approach and there was undercoverage in the perturbed histogram when ϵ was

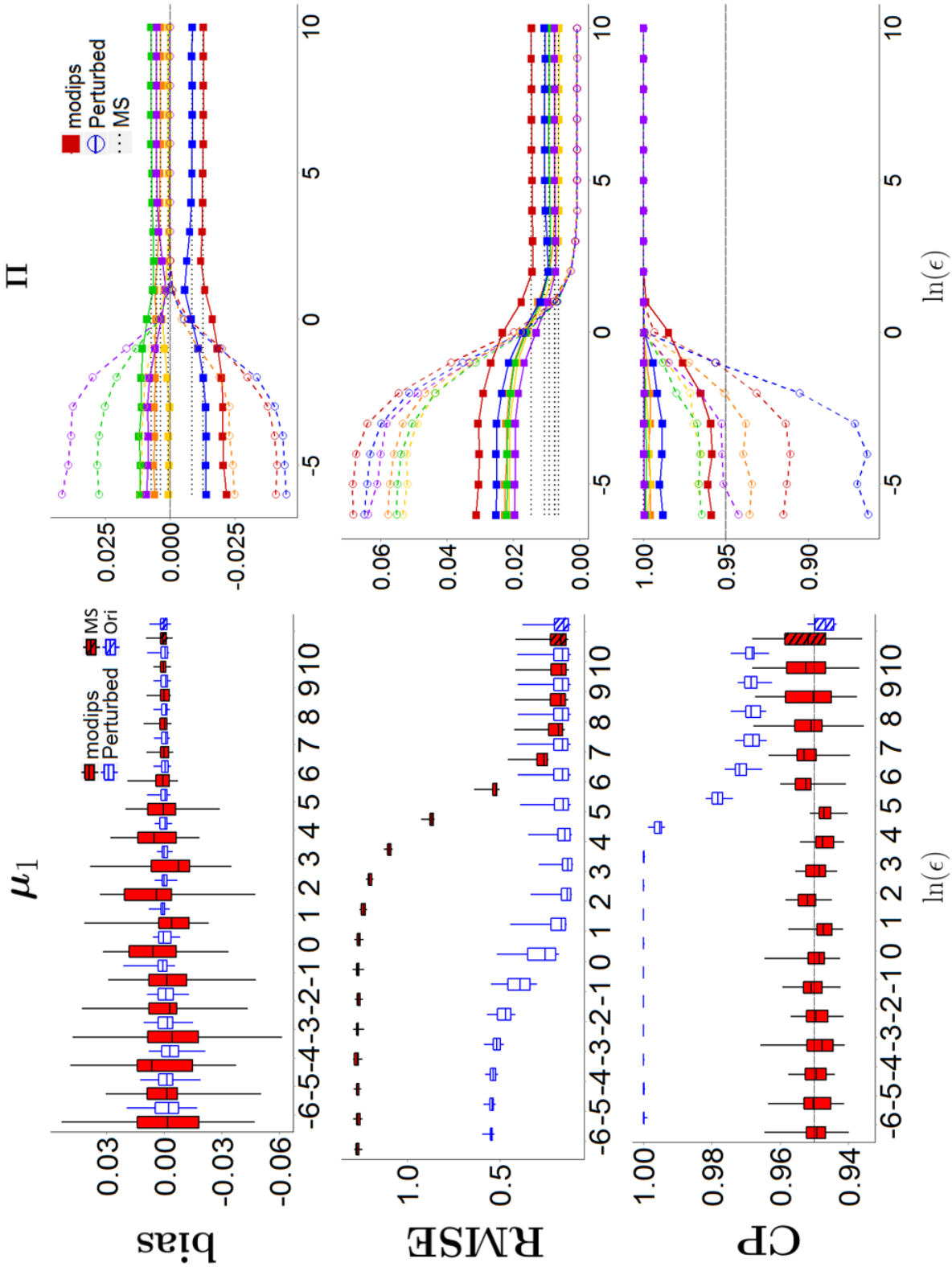


Figure 6: Bias, RMSE, and coverage probability (CP) of μ_1 and II . modips represents the differentially private synthesis with thresholding, perturbed represents the perturbed histogram method, and smoothed represents the smoothed histogram method, Ori is the original results without any perturbation, and MS is the traditional multiple synthesis method without DP.

small. When ϵ was large, the modips technique converged quickly to the MS approach, and while the perturbed histogram converged to the original results. For the inferences on $\boldsymbol{\mu}_1$, the perturbed histogram had more precise estimates than the modips approach. The modips approach still offered valid inferences in terms of nominal-level CP, and small bias; but the perturbed histogram had even smaller bias, smaller RMSE while the CP was at or above 95%. The results were similar to those of $\boldsymbol{\mu}_2$ for both the modips method and the perturbed histogram method. The modips technique outperformed the perturbed histogram in Σ , with much smaller biases and RMSE for all 3 components in $\Sigma(\sigma_1^2, \sigma_2^2, \rho)$. In terms of CP for ρ , the perturbed histogram experienced severe undercoverage in all 3 components in $\sigma_1^2, \sigma_2^2, \rho$ and never reached the nominal level 95%, while the modips approach delivered nominal CP for σ_1^2, σ_2^2 at all ϵ values, and approach the nominal level after $\epsilon > 1$ for ρ . The severe undercoverage in the perturbed histogram even when ϵ was large is not due to the noise added through the Laplace mechanism, but rather because of the discretization in forming the histogram bins. In general, the smaller the number of bins, the larger the variance of the synthetic data is sampled from the piecewise uniform distribution.

5 Discussion

We compared several dips methods conceptually and evaluated the inferential properties of the synthetic data obtained from the dips approaches via extensive simulation studies. These studies were benchmarked against the results from the original data sets without perturbation and the traditional MS approach without DP. From a privacy protection perspective, we recommend the dips techniques over the traditional MS approach since the former provides strong guarantee on privacy protection. All the dips approaches were theoretically proven to satisfy differential privacy at a pre-specified privacy budget when first proposed. Inferences based in the synthetic data from the dips approach are less

precise compared to the traditional MS approach due to the extra layer of noises injected to ensure differential privacy.

The currently available dips methods can be roughly grouped as the nonparametric approaches (np-dips) and the model-based approach (modips). Release of multiple synthetic data sets is necessary when each synthetic data set is generated from a different distribution, such as in the modips approach and the perturbed histogram approach, to account for the uncertainty of a synthesis model. If the synthetic data is generated from a distribution with a fixed set of parameters, then multiple releases are not required. Based on the methodological comparisons and the simulation results, we recommend the modips approach over the other dips approaches for releasing synthetic data. There are a couple of reasons leading this recommendation. First, the modips synthesizer is general and works regardless of data type so long as a proper Bayesian model can be built and Bayesian sufficient statistics can be identified for a given data set and chosen priors. Second, the simulation results showed the modips methods outperformed the other dips methods inferentially overall. In some cases, inferences from the modips approach are less precise (larger RMSE and wider CIs), but it offers nominal coverage probabilities. Since the modips approach is model-based, the inferences would be incorrect if the model is misspecified on a data set. The inferences from the histogram-base approaches are affected by how the histogram bins are formed, an issue independent of DP. One common issue across all the examined dips is the large bias when ϵ is small and the bounds of statistics are not symmetric around the original values. Future work can be devoted to correct for the bias without compromising DP.

Another direction for future work is to apply the dips approaches in real-life data. The most challenging component of the practical application is the large size and complexity of real-life data sets. Real-life data often involves a large number of attributes of various types, not to mention missing data, sparse

data, data entry errors, among others. A regular dips approach without any modification might not accommodate the challenges posed by real-life data. One possible solution in a data set of large n and large p is the data subsetting approach [Liu, 2016a]. Other approaches to improving inferential accuracy in large data is data partitioning [Dwork and Smith, 2010] and application of softer versions of DP (such as (ϵ, δ) probabilistic DP). The only real-life application of a dips approach we could locate is the OnTheMap project [Machanavajjhala et al., 2008], which releases worker commuter patterns data collected by US Census Bureau. The direct application of the MD-synthesizer led to poor statistical inferences due to the sparsity of the data on a large domain (8.2 million Census blocks - origin location). To address this issue, the authors combined distance-based coarsening with a probabilistic pruning algorithm [Abowd and Vilhuber, 2008].

In summary, to the best of our knowledge, this paper is the first work that compares the utility and inferential properties of the synthetic data released via the np-dips and modips approaches. We expect our findings to shed light on the data utility aspect as well as the feasibility of applying the dips approach in data sets of various types and sizes, enabling us to move one step closer to achieving the goal of releasing complex differentially private synthetic data for public use.

References

- A. Narayanan and V. Shmatikov. Robust de anonymization of large sparse datasets. *IEEE Symposium on Security and Privacy*, pages 111–125, 2008.
- N Homer, S Szelling, M Redmann, D Duggan, W Tembe, J Muehling, JV Pearson, DA Stephan, SF Nelson, and DW Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genet*, 4(8):e1000167, 2008.
- M. Götz, A. Machanavajjhala, G. Wang, X. Xiao, and J. Gehrke. Publishing search logs - a comparative study of privacy guarantees. *IEEE Trans. Knowl. Data Eng.*, 24:5205325, 2012.
- Latanya Sweeney. Matching known patients to health records in washington state data. *Available at SSRN 2289850*, 2013.
- J. Drechsler. *Synthetic datasets for Statistical Disclosure Control*. Springer, New York, 2011.
- Trivellore E Raghunathan, Jerome P Reiter, and Donald B Rubin. Multiple imputation for statistical disclosure limitation. *Journal of official Statistics*, 19(1):1–16, 2003.
- Jerome P Reiter. Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, 18(4):531–543, 2002.
- Jerome P Reiter. Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29(2):181–188, 2003.
- Jerome P Reiter. Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association*, 100(472):1103–1112, 2005.

- Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, and Peter-Paul De Wolf. *Statistical disclosure control*. John Wiley & Sons, 2012.
- Daniel Manrique-Vallier and Jerome P Reiter. Estimating identification disclosure risk using mixed membership models. *Journal of the American Statistical Association*, 107(500):1385–1394, 2012.
- John M Abowd and Lars Vilhuber. How protective are synthetic data? In *Privacy in Statistical Databases*, pages 239–246. Springer, 2008.
- Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography*, pages 265–284. Springer, 2006.
- C. Dwork. Differential privacy: A survey of results. *Theory and Applications of Models of Computation*, 4978:1–19, 2008.
- Cynthia Dwork. Differential privacy. In *Encyclopedia of Cryptography and Security*, pages 338–340. Springer, 2011.
- Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Foundations of Computer Science, 2007. FOCS’07. 48th Annual IEEE Symposium on*, pages 94–103. IEEE, 2007.
- Frank McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 19–30. ACM, 2009.
- Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *Proceedings*

of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pages 273–282. ACM, 2007.

Cynthia Dwork and Adam Smith. Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, 1(2):2, 2010.

Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 609–618. ACM, 2008.

Noman Mohammed, Rui Chen, Benjamin Fung, and Philip S Yu. Differentially private data release for data mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 493–501. ACM, 2011.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *JMLR: Workshop and Conference Proceedings*, 12:1069–1109, 2011.

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. *JMLR: Workshop and Conference Proceedings*, 23:25.1–25.40, 2012.

Kamalika Chaudhuri, Anand Sarwate, and Kaushik Sinha. Near-optimal differentially private principal components. *Proc. 26th Annual Conference on Neural Information Processing Systems (NIPS)*, 2012.

Fei Yu, Stephen E. Fienberg, Aleksandra B. Slavkovic, and Caroline Uhler. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of Biomedical Informatics*, 50: 133–141, 2014.

- A. S. Charest. How can we analyze differentially private synthetic datasets. *Journal of Privacy and Confidentiality*, 2(2):Article 3, 2010.
- David McClure and Jerome P Reiter. Differential privacy and statistical disclosure risk measures: An investigation with binary synthetic data. *Transactions on Data Privacy*, 5(3):535–552, 2012.
- Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.
- F. Liu. Model-based differential private data synthesis. *in preparation*, 2016a.
- Rathindra Sarathy and Krish Muralidhar. Differential privacy for numeric data. *Joint UN-ECE/Eurostat work session on statistical data confidentiality, Bilbao, Spain*, 2009.
- F. Liu. Some notes on laplace mechanism and exponential mechanism in differential privacy. *in preparation*, 2016b.
- Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 9(3-4):211–407, 2013.
- David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.
- Jouni Kerman. Neutral noninformative and informative conjugate beta and gamma prior distributions. *Electronic Journal of Statistics*, 5:1450–1470, 2011.
- Ashwin Machanavajjhala, Daniel Kifer, John Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets practice on the map. *IEEE ICDE IEEE 24th International Conference*, pages 277 – 286, 2008.

6 Supplemental Materials

This section contains the supplementary materials to accompany the paper “Differential Private Data Synthesis Methods” with additional results from the three simulation studies. Table 1 shows the proportion of usable simulation repeats for all methods in the first simulation study. Figure 1 depicts the results when $\pi = 0.50$ and Figure 2 shows the average width of the 95% CI of π in the first simulation study. Table 2 contains the summary statistics for the number of histogram bins, and Figure 3 depicts the average average width of the 95% CI of μ from the second simulation study, when the bounds of the data were $[\mu - 3\sigma, \mu + 4\sigma]$. Figures 4 to 7 show the results of the parameters μ and σ^2 from the second simulation, when the bounds of the data were $[\mu - 4\sigma, \mu + 4\sigma]$. Figures 8 and 9 depict the results of the parameter σ^2 from the second simulation, when the bounds of the data were $[\mu - 3\sigma, \mu + 4\sigma]$. Table 3 contains the summary statistics of μ_1 , μ_2 , and π for the 24 cells, Table 4 contains the summary statistics for the number of 2-dimensional histogram bins for the continuous variables $(\mathbf{z}_1, \mathbf{z}_2)$, and Figures 10 and 11 show the results of the parameters μ_2 , ρ , σ_1^2 , and σ_2^2 from the third simulation. Table 5 contains the average frequency of empty cells in a synthetic data set from the third simulation study.

Table 1: The proportion of usable simulation repeats (out of 5000) in calculation of 95% CI for all methods in the first simulation study with categorical data. A “usable” simulation repeat is defined as a simulated data set that leads to a synthetic data set that contains at least one of each of the two levels of the binary variable

The proportion of usable repeats for the Laplace Sanitizer method

$\ln(\epsilon)$	-10	-9	-8	-7	-6	-5
$n = 40, \pi = 0.15$	0.0008	0.0018	0.0074	0.0190	0.0448	0.1140
$n = 100, \pi = 0.15$	0.0016	0.0074	0.0140	0.0446	0.1144	0.2672
$n = 1000, \pi = 0.15$	0.0240	0.0548	0.1480	0.3160	0.5942	0.8078
$n = 40, \pi = 0.25$	0.0008	0.0020	0.0078	0.0174	0.0472	0.1214
$n = 100, \pi = 0.25$	0.0020	0.0076	0.0158	0.0458	0.1120	0.2748
$n = 1000, \pi = 0.25$	0.0238	0.0568	0.1510	0.3310	0.6496	0.8990
$n = 40, \pi = 0.50$	0.0012	0.0024	0.0074	0.0162	0.0464	0.1336
$n = 100, \pi = 0.50$	0.0024	0.0070	0.0190	0.0458	0.1152	0.2868
$n = 1000, \pi = 0.50$	0.0240	0.0584	0.1554	0.3558	0.7058	0.9658

The proportion of usable repeats for the modips, MD and DP-prior approaches

	modips-piecewise		modips-truncated		Multinomial-Dirichlet		DP-Prior	
$\ln(\epsilon)$	-10	-9	-10	-9	-10	-9	-10	-9
$n = 40, \pi = 0.15$	0.9996	0.9994	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$n = 100, \pi = 0.15$	1.0000	0.9996	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$n = 1000, \pi = 0.15$	0.9998	0.9994	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$n = 40, \pi = 0.25$	0.9998	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$n = 100, \pi = 0.25$	0.9998	0.9996	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$n = 1000, \pi = 0.25$	0.9998	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$n = 40, \pi = 0.50$	0.9994	0.9996	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$n = 100, \pi = 0.50$	0.9994	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$n = 1000, \pi = 0.50$	0.9998	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

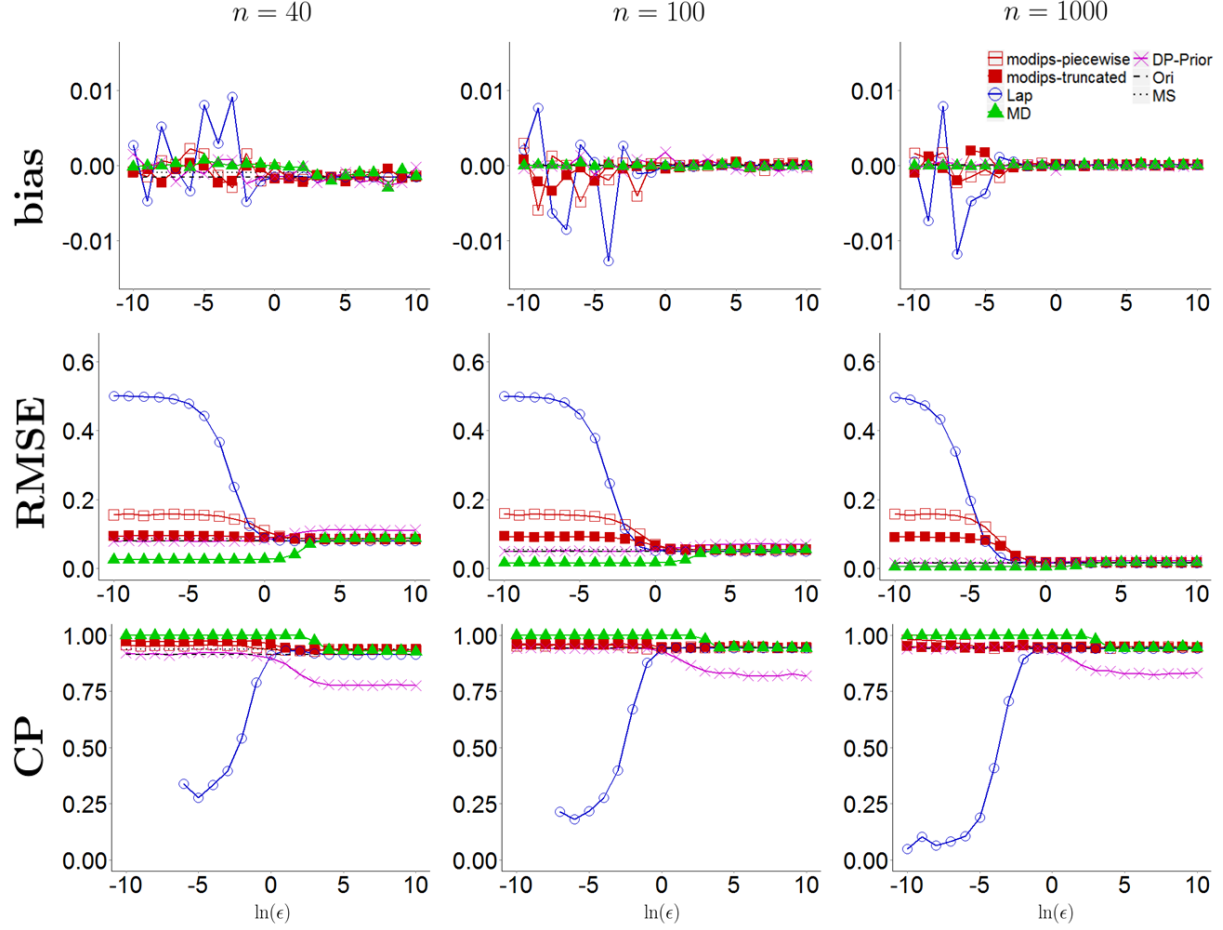


Figure 1: Bias, RMSE, and CP of $\pi = 0.50$. modips represents the model-based differentially private synthesis, Lap represents the Laplace synthesizer, MD represents the MD synthesizer, DP-prior represents DP-prior synthesizer, Ori is the original results without any perturbation, and MS is the traditional multiple synthesis method without DP.

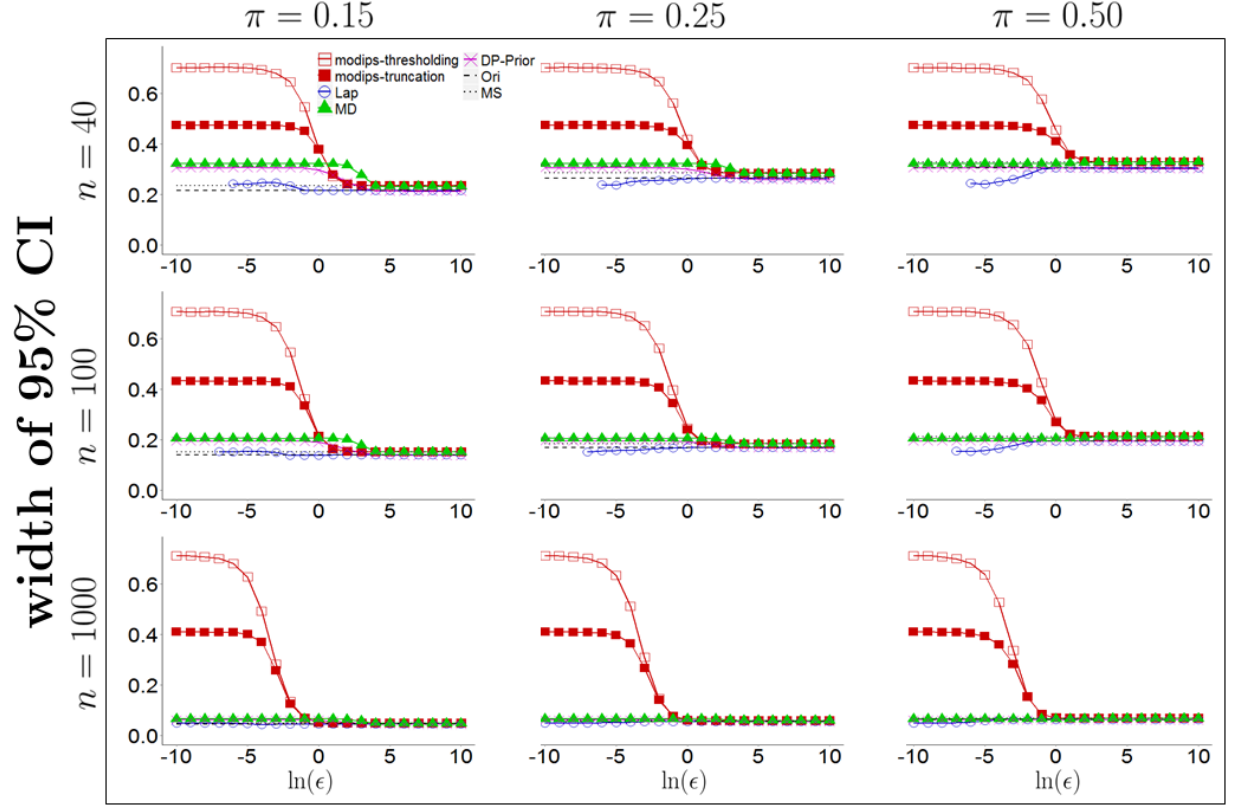


Figure 2: The average width of the 95% CI of π . *modips* represents the model-based differentially private synthesis, *Lap* represents the Laplace synthesizer, *MD* represents the MD synthesizer, *DP-prior* represents DP-prior synthesizer, *Ori* is the original results without any perturbation, and *MS* is the traditional multiple synthesis method without DP.

Table 2: Summary statistics for the number of bins in the second simulation study with continuous data

Scenario	Min	Mean	Median	Max
$n = 40, \sigma^2 = 1, 4, 9$	5	7.459	7	12
$n = 100, \sigma^2 = 1, 4, 9$	8	9.853	10	13
$n = 1000, \sigma^2 = 1, 4, 9$	19	20.54	21	22

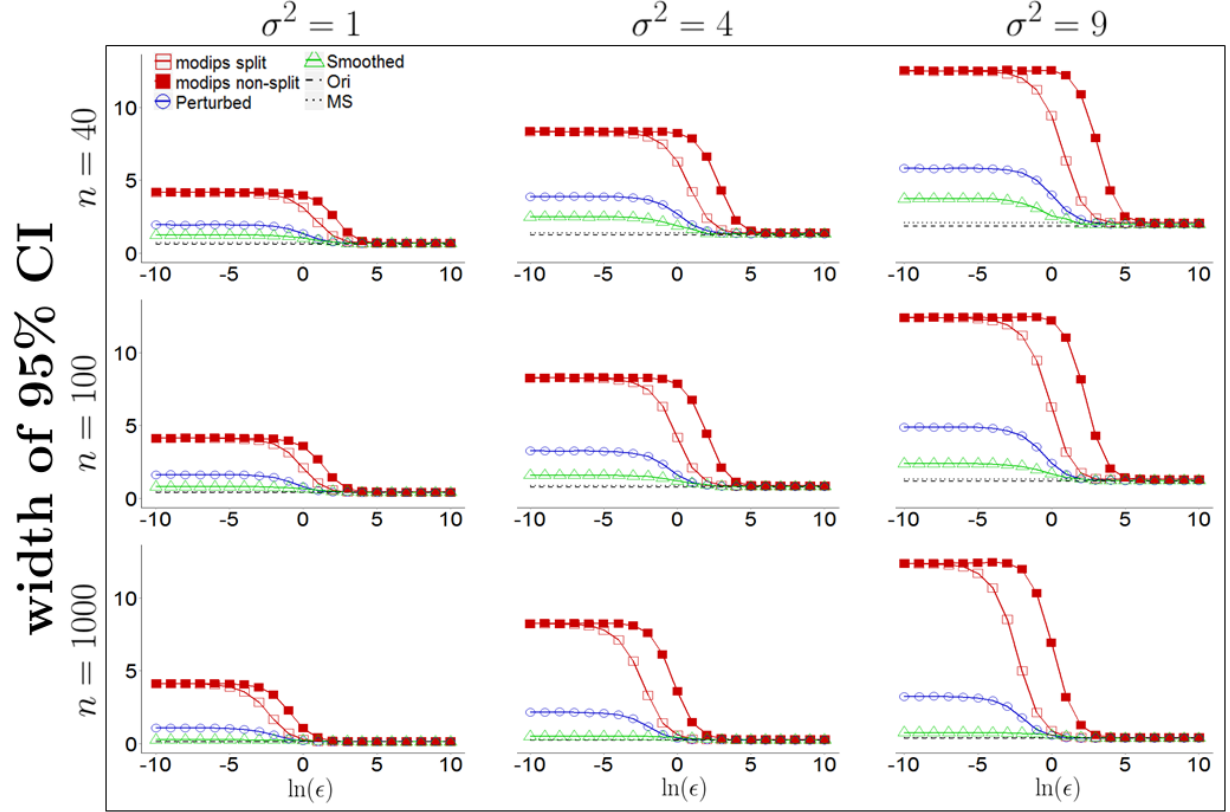


Figure 3: The average width of the 95% CI of μ from the second simulation, when the bounds are $[\mu - 3\sigma, \mu + 4\sigma]$. *modips-split* and *modips-nonsplit* represent the model-based differentially private synthesis (with thresholding) with ϵ split and not split (respectively) between μ and σ^2 , *perturbed* represents the perturbed histogram method, *smoothed* represents the smoothed histogram method, *Ori* is the original results without any perturbation, and *MS* is the traditional multiple synthesis method without DP.

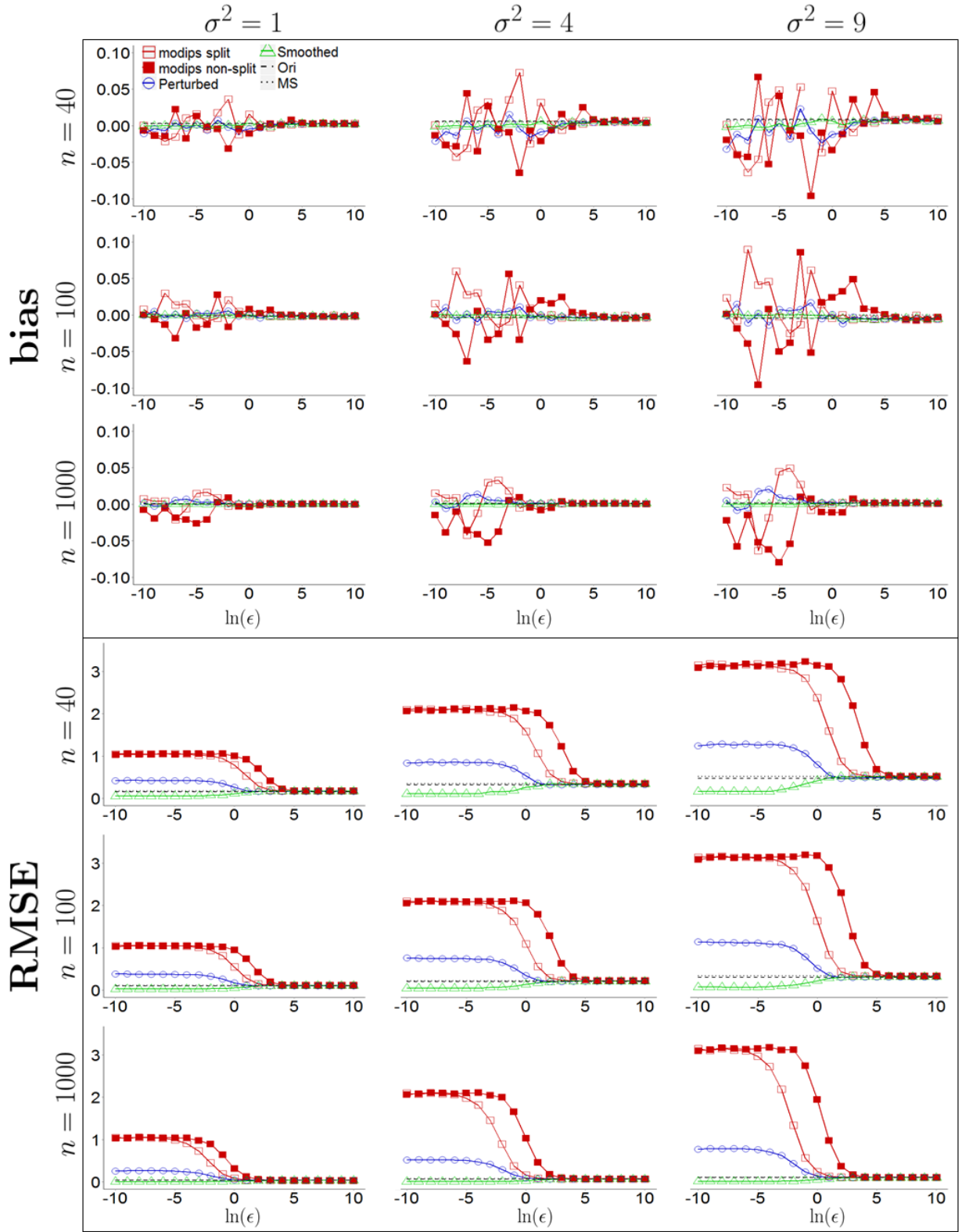


Figure 4: The bias and RMSE of μ from the second simulation, when the bounds are $[\mu - 4\sigma, \mu + 4\sigma]$. *modips-split* and *modips-nonsplit* represent the model-based differentially private synthesis (with thresholding) with ϵ split and not split (respectively) between μ and σ^2 , *perturbed* represents the perturbed histogram method, *smoothed* represents the smoothed histogram method, *Ori* is the original results without any perturbation, and *MS* is the traditional multiple synthesis method without DP.

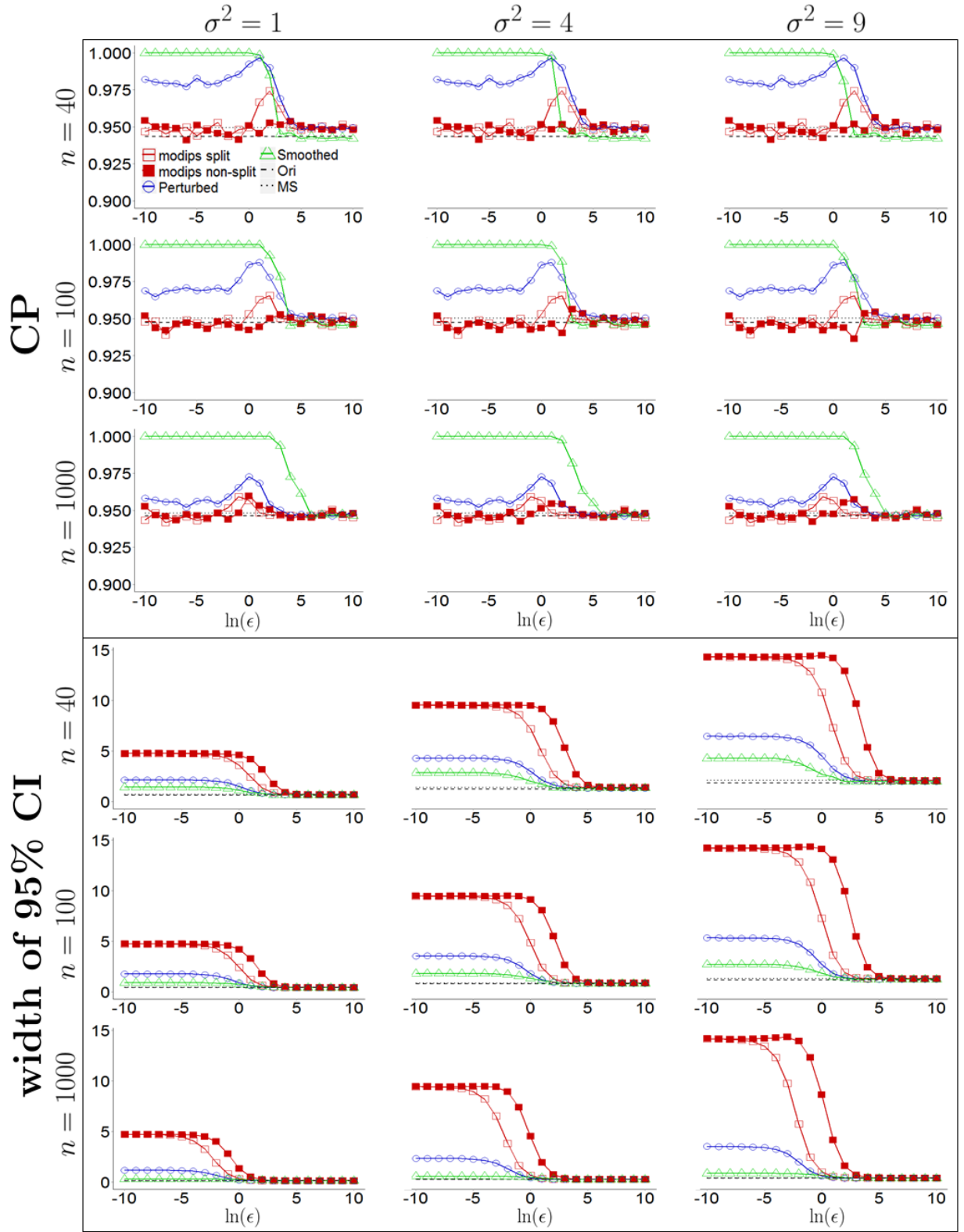


Figure 5: The CP and average width of 95% CI of μ from the second simulation, when the bounds were $[\mu - 4\sigma, \mu + 4\sigma]$. modips-split and modips-nonsplit represent the model-based differentially private synthesis (with thresholding) with ϵ split and not split (respectively) between μ and σ^2 , perturbed represents the perturbed histogram method, smoothed represents the smoothed histogram method, Ori is the original results without any perturbation, and MS is the traditional multiple synthesis method without DP.

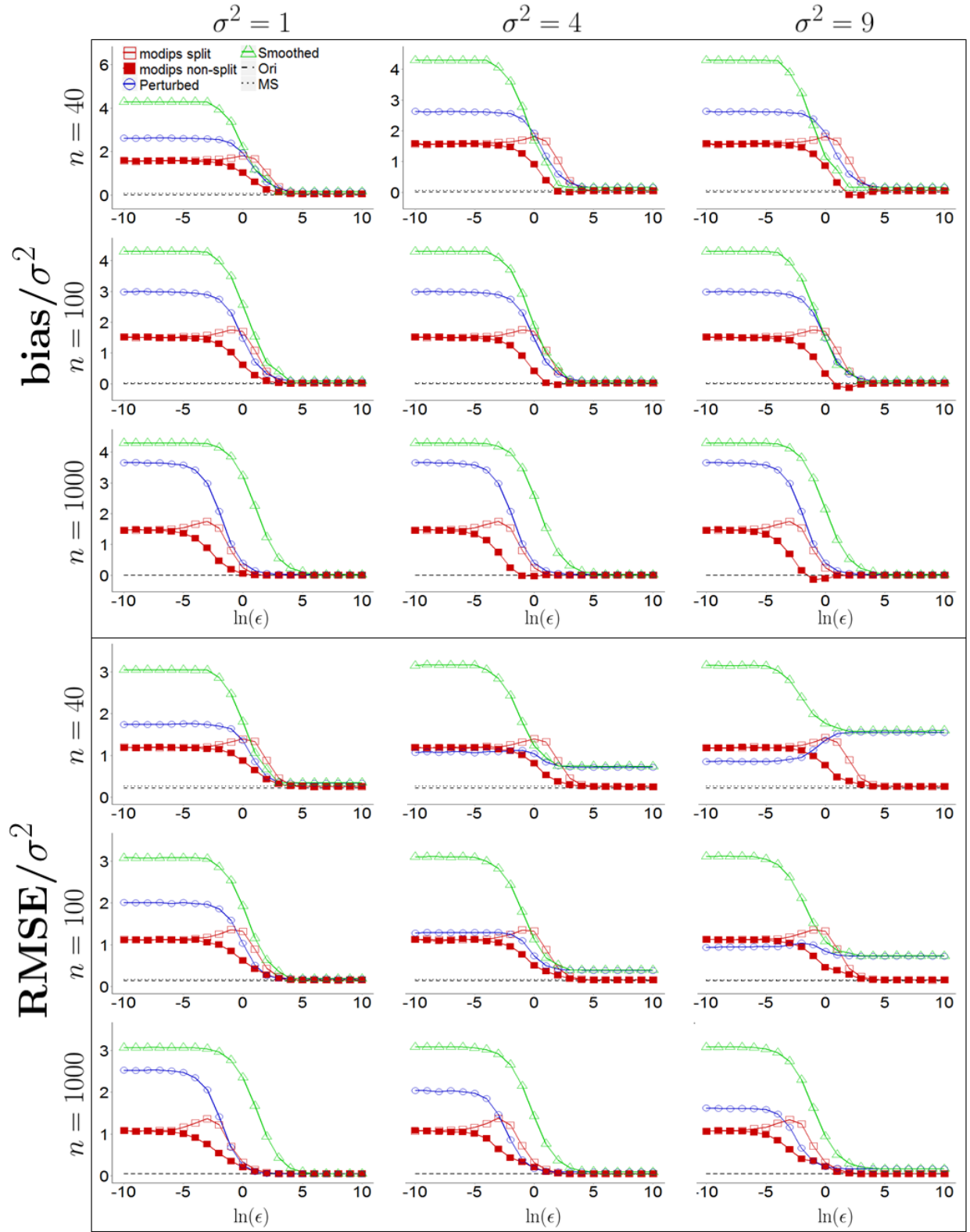


Figure 6: The relative bias and RMSE/σ^2 of σ^2 from the second simulation, when the bounds were $[\mu - 4\sigma, \mu + 4\sigma]$. modips-split and modips-nonsplit represent the model-based differentially private synthesis (with thresholding) with ϵ split and not split (respectively) between μ and σ^2 , perturbed represents the perturbed histogram method, smoothed represents the smoothed histogram method, Ori is the original results without any perturbation, and MS is the traditional multiple synthesis method without DP.

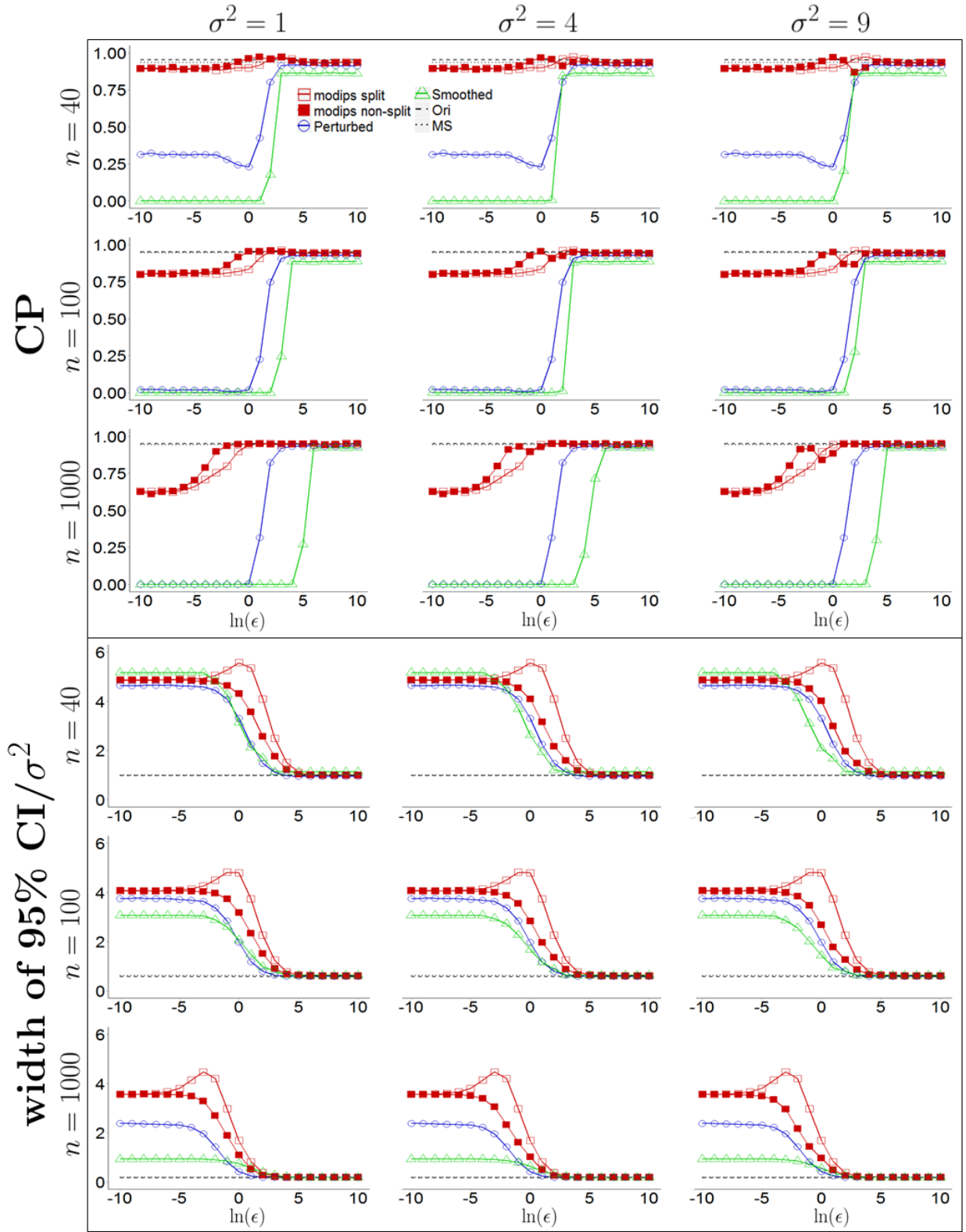


Figure 7: The CP and average width of 95% CI/ σ^2 of σ^2 from the second simulation, when the bounds were $[\mu - 4\sigma, \mu + 4\sigma]$. modips-split and modips-nonsplit represent the model-based differentially private synthesis (with thresholding) with ϵ split and not split (respectively) between μ and σ^2 , perturbed represents the perturbed histogram method, smoothed represents the smoothed histogram method, Ori is the original results without any perturbation, and MS is the traditional multiple synthesis method without DP.

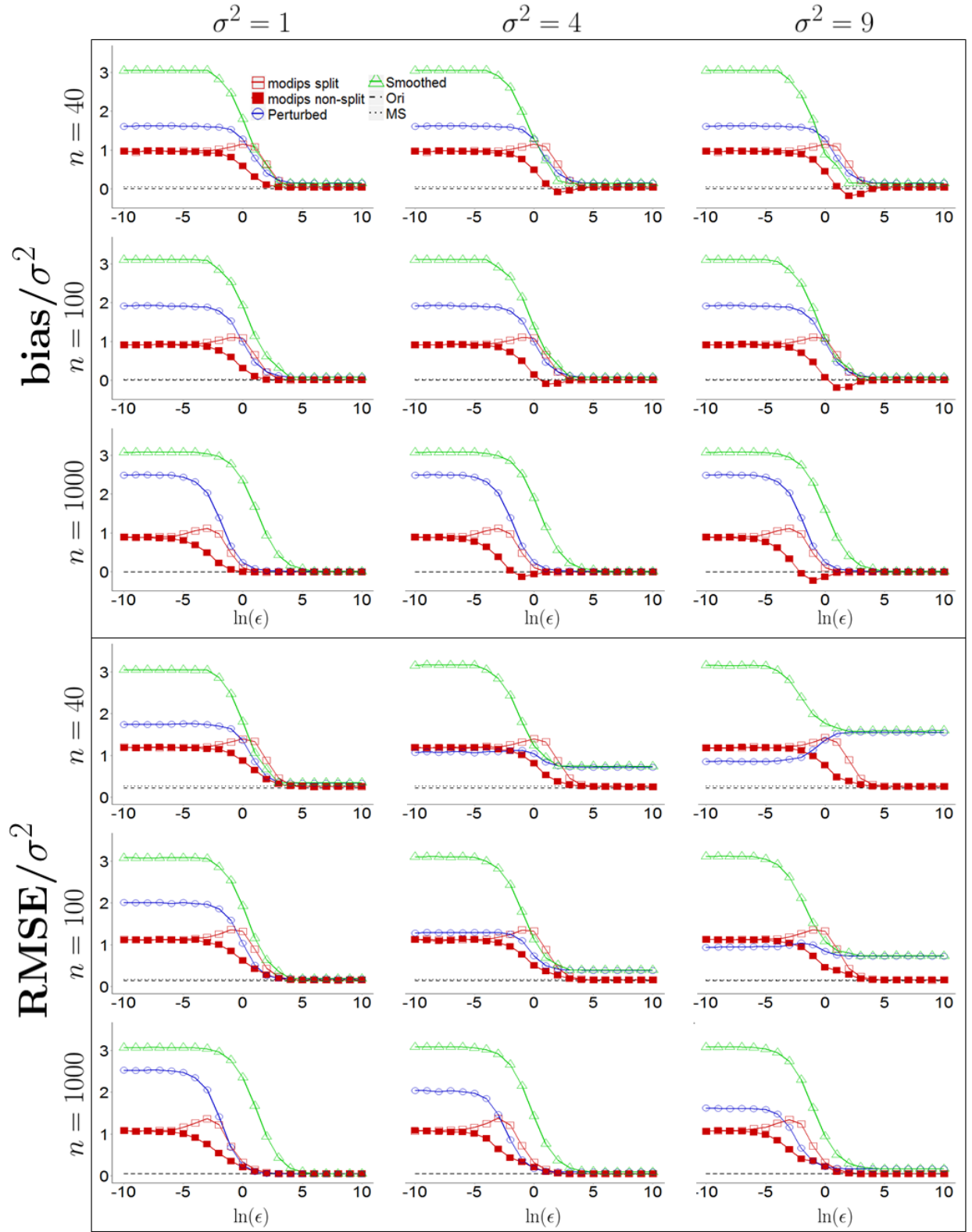


Figure 8: The relative bias and RMSE/σ^2 from the second simulation, when the bounds were $[\mu - 3\sigma, \mu + 4\sigma]$. *modips-split* and *modips-nonsplit* represent the model-based differentially private synthesis (with thresholding) with ϵ split and not split (respectively) between μ and σ^2 , *perturbed* represents the perturbed histogram method, *smoothed* represents the smoothed histogram method, *Ori* is the original results without any perturbation, and *MS* is the traditional multiple synthesis method without DP.

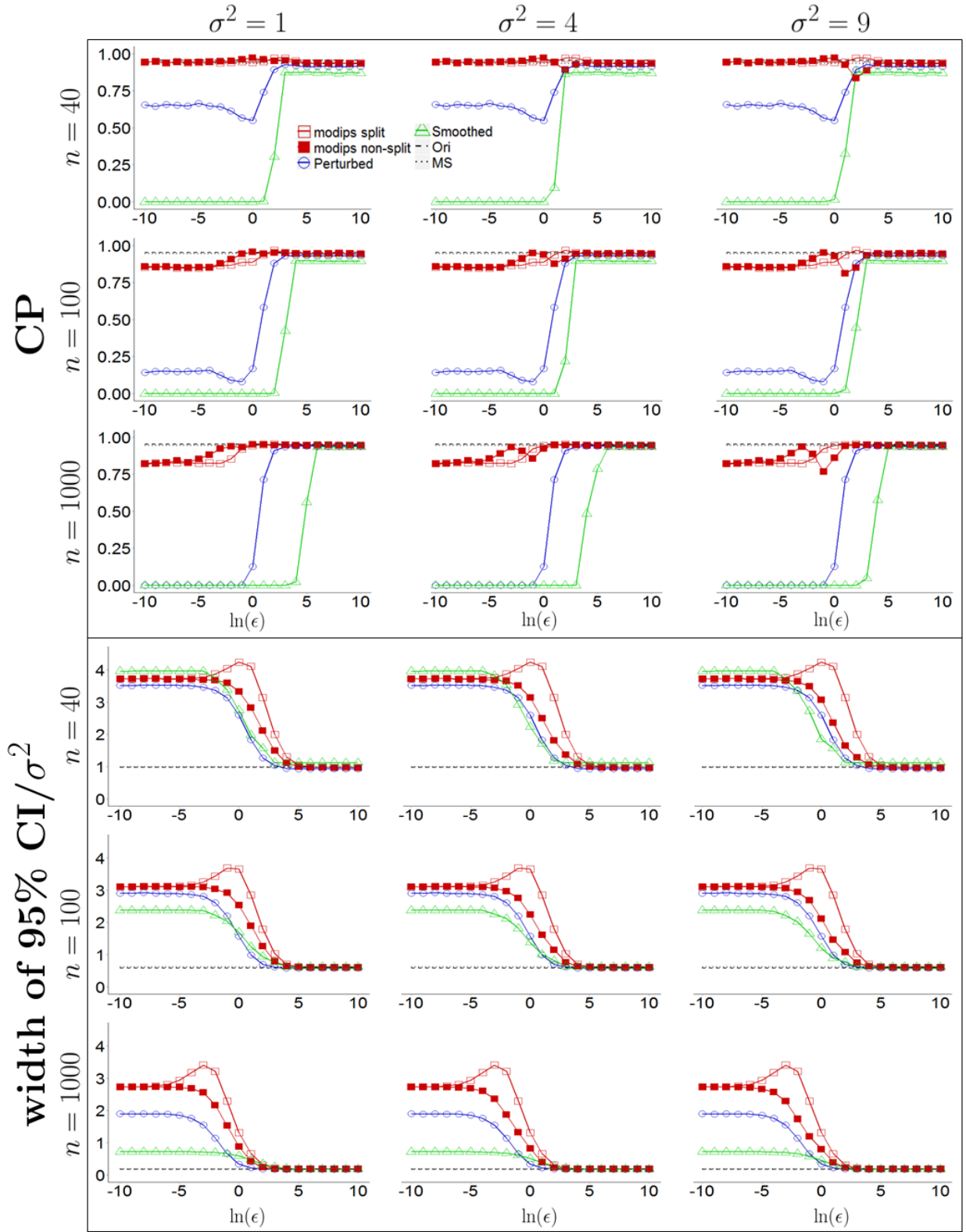


Figure 9: The CP and average width of 95% CI/ σ^2 of σ^2 from the second simulation, when the bounds were $[\mu - 3\sigma, \mu + 4\sigma]$. modips-split and modips-nonsplit represent the model-based differentially private synthesis (with thresholding) with ϵ split and not split (respectively) between μ and σ^2 , perturbed represents the perturbed histogram method, smoothed represents the smoothed histogram method, Ori is the original results without any perturbation, and MS is the traditional multiple synthesis method without DP.

Table 3: Summary statistics of μ_1 , μ_2 , and π for the 24 cells in the third simulation study—mixture model simulation study.

Parameter	Min	Q1	Median	Q3	Max
μ_1	-2.440	-0.426	0.285	0.827	2.018
μ_2	-2.657	-0.662	-0.193	0.634	2.287
π	0.007	0.0240	0.043	0.0620	0.0760

Table 4: Summary statistics for the number of 2-dimensional histogram bins with the two continuous variables in the third simulation study

Cell	Min	Mean	Median	Max
1	36	37.78	36	49
2	49	51.09	49	64
3	25	34.39	36	36
4	42	49.01	49	56
5	36	42.06	42	49
6	36	37.78	36	49
7	36	36.38	36	49
8	16	16.04	16	20
9	49	49.04	49	56
10	49	49.04	49	56
11	36	36.00	36	36
12	36	45.60	49	49
13	36	48.49	49	49
14	16	16.04	16	20
15	25	28.59	30	36
16	49	49.06	49	56
17	49	49.37	49	64
18	20	25.00	25	25
19	25	34.39	36	36
20	25	35.97	36	36
21	36	45.60	49	49
22	16	24.92	25	25
23	36	48.97	49	56
24	42	49.01	49	56

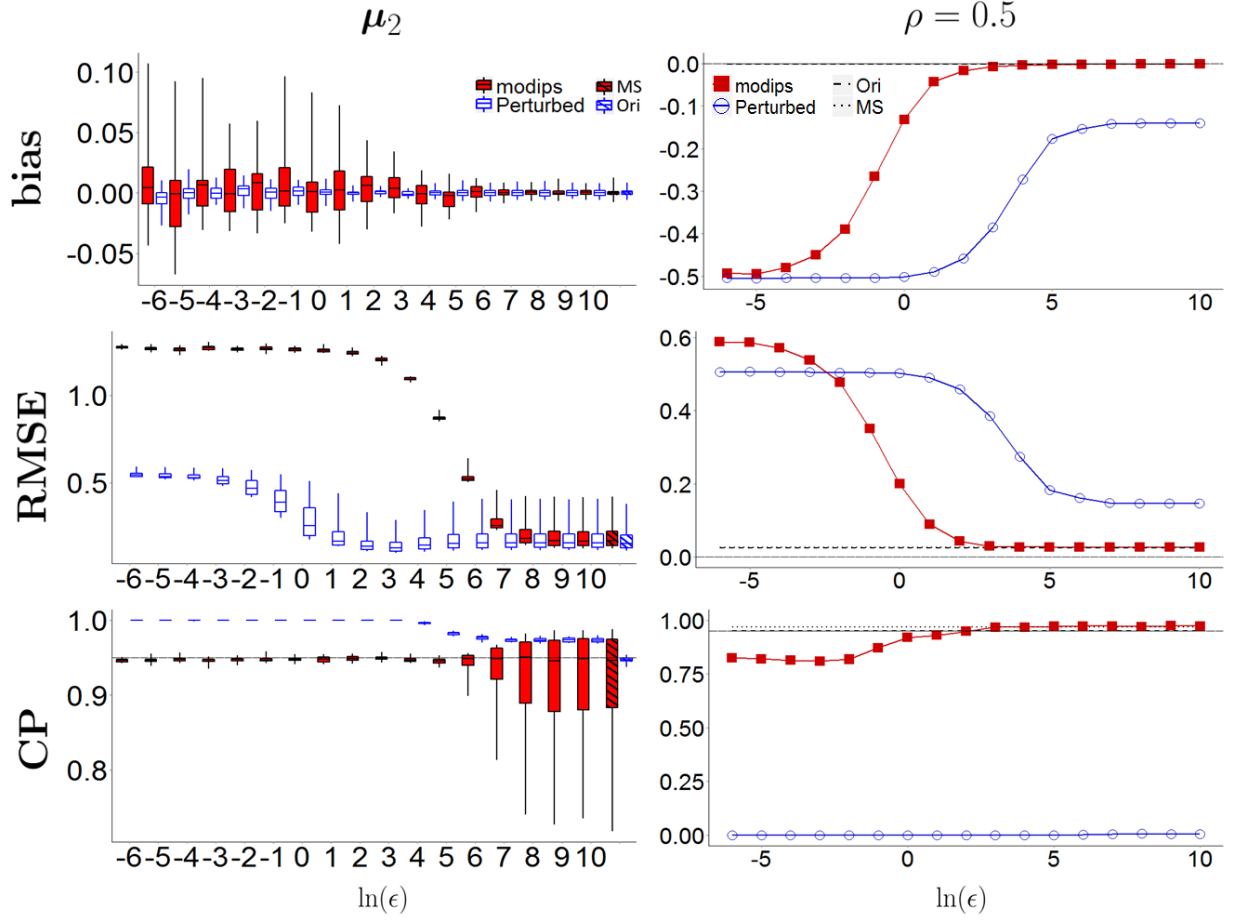


Figure 10: The bias, RMSE, and CP of μ_2 and ρ from the third simulation. *modips* represents the differentially private synthesis (with thresholding), *perturbed* represents the perturbed histogram method, *Ori* is the original results without any perturbation, and *MS* is the traditional multiple synthesis method without DP.

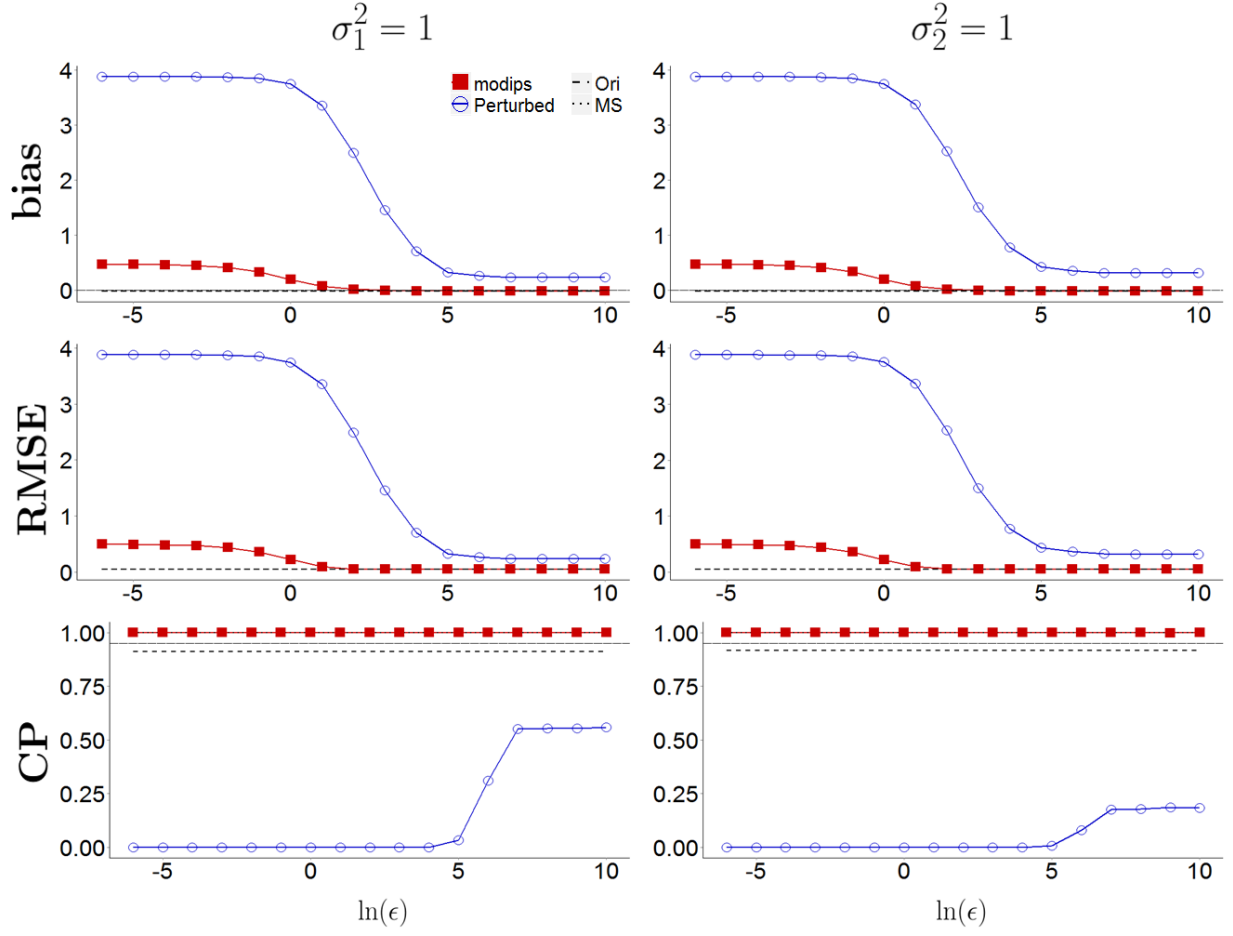


Figure 11: The bias, RMSE, and CP of σ_1^2 , and σ_2^2 from the third simulation. *modips* represents the differentially private synthesis (with thresholding), *perturbed* represents the perturbed histogram method, *Ori* is the original results without any perturbation, and *MS* is the traditional multiple synthesis method without DP.

Table 5: The average frequency of empty cells (i.e. when $n_k^* = 0$ for $k = 1, \dots, 24$) per synthetic data set by the modips (with thresholding) and the perturbed histogram methods

$\ln(\epsilon)$	modips	Perturbed Histogram
-6	3.81	23.84
-5	3.78	23.82
-4	3.66	23.80
-3	3.72	23.75
-2	3.82	23.52
-1	3.88	20.62
0	4.02	8.95
1	4.09	4.43
2	3.49	2.09
3	2.41	0.19
4	0.99	0.00
5	0.70	0.00
6	0.76	0.00
7	0.70	0.00
8	0.76	0.00
9	0.65	0.00
10	0.68	0.00